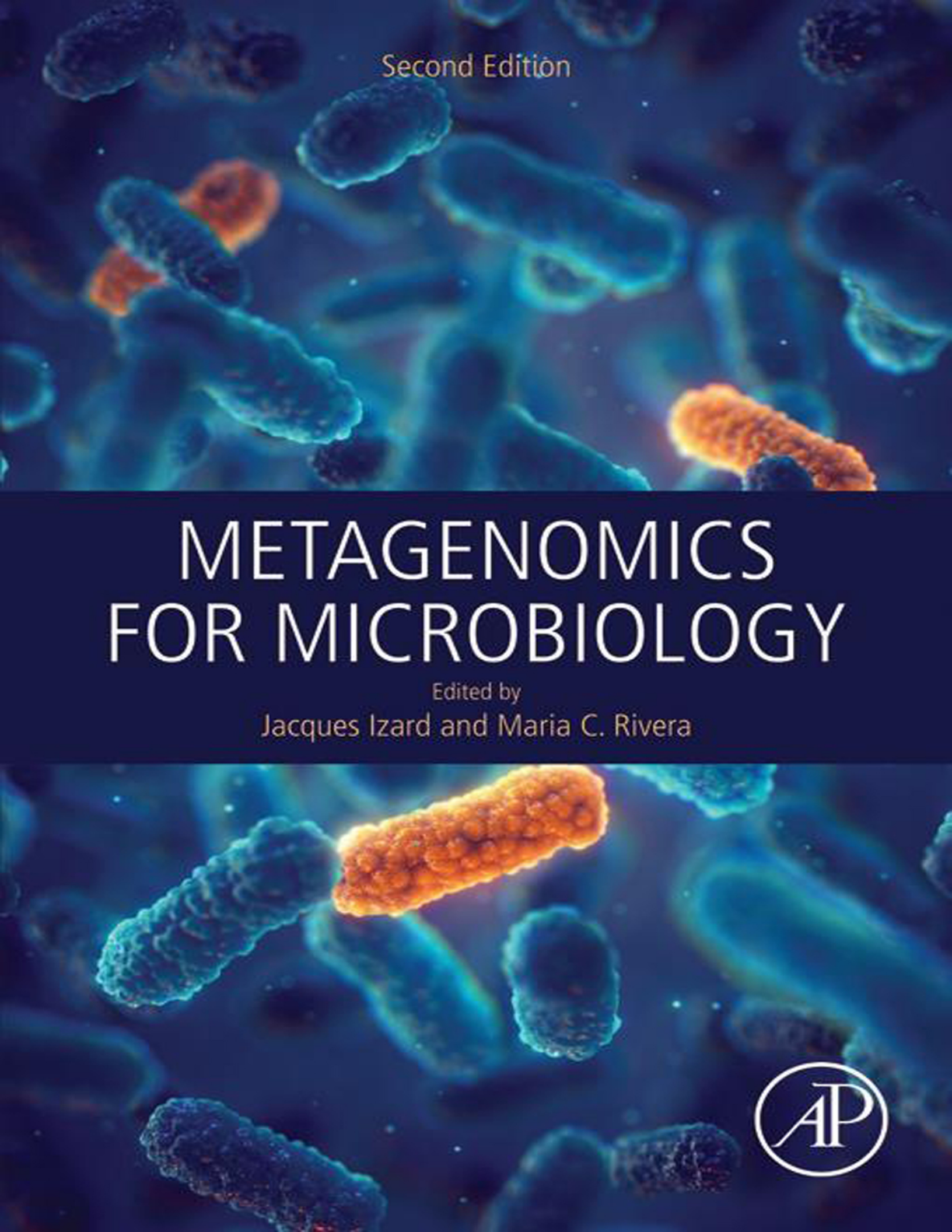

Second Edition
METAGENOMICS
FOR MICROBIOLOGY
Edited by
Jacques Izard and Maria C. Rivera

Metagenomics for Microbiology

SECOND EDITION

EDITED BY

Jacques Izard

Department of Internal Medicine, University of Nebraska Medical Center, Omaha, NE, United States

Maria C. Rivera

School of Life Sciences and Sustainability, Virginia Commonwealth University, Richmond, VA, United States



ACADEMIC PRESS

An imprint of Elsevier

Table of Contents

Cover image
Title page
Copyright
List of contributors
Preface

Chapter 1. Toward success in omics studies
- **Abstract**
- **Metagenomics and why we consider it a pipeline**
- **Study design: aligning your steps to improve overall quality**
- **Statistical analysis plan: an early step, not an after-thought**
- **Ethical and legal issues**
- **The so-called “contamination by the host and/or environment”**
- **Sample acquisition and storage: key steps to get the best data**
- **Sample extraction: a step closer to high-quality datasets**
- **Sequencing: an equation between hypothesis, sequencing depth, and financial matters**
- **Metadata: a context for the analysis and the door to our overall understanding**
- **Data analysis: finally getting the information you have been waiting for**

Details frequently overlooked (snippets)
Data publication and dissemination
Integrating the scales
Examining the status quo
Acknowledgments
References
Chapter 2. The changing landscape of microbiome characterization methods: tools, trends, and trajectories
Abstract
The way forward: traditional technologies scaled up and new -omics approaches
The next decade sees shifting methodologies and an increased emphasis on large datasets
Looking forward: expectations and predictions
Conclusions
Acknowledgments
References
Chapter 3. The use of metagenomics for biothreat and emerging threat applications
Abstract
Introduction
Clinical utility
Food and drug safety
Fieldwork
Surveillance
Epidemiological investigation
Climate change and emerging threats
Metagenomic next-generation sequencing approaches
Availability of commercial kits/instruments
Laboratory developed workflows
Sample processing

Sequence analysis
Challenges
Dual use potential and cyberbiosecurity
Conclusion
References
Chapter 4. Metagenomics, microbiomes, and symbioses
Abstract
Introduction
Endosymbioses in the origin of eukaryotes and mitochondria
Endosymbioses and plastids evolution
Symbioses in protists
Endosymbionts with potential for denitrification and aerobic respiration
Host genotype structures the microbiome of a diatom
Choanoflagellates microbiome and the origin of metazoans
Microbiomes and symbioses across the invertebrates
Helminths: microbiomes within microbiomes
Conclusions
References
Chapter 5. Hypothesis testing of metagenomic data
Abstract
Introduction
Metagenomic data
Compare diversity across groups
Comparing a single taxon across groups
Compare all taxa across groups simultaneously
Discussion and future research
References
Chapter 6. An update and review of ribosomal RNA depletion methods for microbial transcriptomics

Abstract
Introduction
Size selection by gel electrophoresis
Denaturing high-performance liquid chromatography
5 ′ phosphate-dependent exonuclease digestion
Subtractive hybridization
RNase H digestion
Clustered regularly interspersed short palindromic repeats/Cas9-based rRNA depletion
Blocking reverse transcription of rRNA in library preparation
Enrichment of mRNA by blocked rRNA
Biased 3′ polyadenylation against rRNA
Selective priming in cDNA synthesis
cDNA library normalization
Combination of methods
Ribosomal RNA depletion in dual RNA-seq
Ribosomal RNA depletion in single-cell RNA-seq
Conclusions and perspectives
References
Chapter 7. Anything, anyone, anywhere: metagenomic sequencing with Oxford Nanopore Technologies
Abstract
Introduction
How does Oxford Nanopore sequencing work?
How does nanopore sequencing enhance metagenomic research?
From sample to answer: the metagenomic Oxford Nanopore sequencing workflow
Metagenomic sequencing or 16S sequencing?

Applications of metagenomic nanopore sequencing for human and environmental health
Conclusion
References
Chapter 8. Highly accurate long-read PacBio HiFi sequencing for microbiome and metagenomic applications
Abstract
Introduction
Full-length gene amplicon microbiome sequencing
Taxonomic and functional profiling of metagenomic reads
Metagenome assembly
Concluding remarks
References
Chapter 9. Genome centric metagenomics: concepts and resources
Abstract
What is a metagenome-assembled genome?
How do you construct a metagenome-assembled genome?
Metagenome-assembled genome quality standards
Naming and organizing the uncultivated
What are key differences between a metagenome-assembled genome and an isolate genome?
Conclusions and outlook
Acknowledgments
References
Chapter 10. Understanding and modeling the structure of omics data through the lens of probability distributions
Abstract
Why we need new methodologies for determining probability distributions in microbiome studies

Men's lifestyle validation study is an excellent resource for studying distributional profiles of microbiome attributes
Intelligent decision support framework methodology
Component 1 of the intelligent decision support framework
Components 2 and 3 of the intelligent decision support framework
Component 4 of the intelligent decision support framework
Rationale for the intelligent decision support framework
Observed distributions of microbiome attributes as identified using the intelligent decision support framework
Application of the intelligent decision support framework to assess diet-microbiome associations
The intelligent decision support framework in context
Acknowledgments
References
Chapter 11. Metabolomics in microbial ecology: sample preservation, storage, and integration with metagenomics
Abstract
Why metabolomics
Targeted versus untargeted metabolomics
Sample handling considerations for metabolomics analyses
Tissue
Dried blood spots
Sample extraction
What are the instruments used?
Data processing and annotation
Integration of metabolomics and metagenomics
Conclusion and future directions
References
Chapter 12. The event of multiomics in metagenomics
Abstract

Introduction
Multiomics approaches in metagenomics
Frameworks for microbiome multiomics integration
Success stories in microbiome multiomics
Conclusions and future directions
References

Glossary
Index

Copyright

Academic Press is an imprint of Elsevier
125 London Wall, London EC2Y 5AS, United Kingdom
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States

Copyright © 2025 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

For accessibility purposes, images in electronic versions of this book are accompanied by alt text descriptions provided by Elsevier. For more information, see **https://www.elsevier.com/about/accessibility**.

Books and Journals published by Elsevier comply with applicable product safety requirements. For any product safety concerns or queries, please contact our authorised representative, Elsevier B.V., at **productsafety@elsevier.com**.

Publisher's note: Elsevier takes a neutral position with respect to territorial disputes or jurisdictional claims in its published content, including in maps and institutional affiliations.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information

about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: **www.elsevier.com/permissions**.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

ISBN: 978-0-443-22376-1

For Information on all Academic Press publications visit our website at **https://www.elsevier.com/books-and-journals**

Publisher: Stacy Masucci
Acquisitions Editor: Patricia Osborn

Editorial Project Manager: Pat Gonzalez
Production Project Manager: Punithavathy Govindaradjane
Cover Designer: Christian Bilbow

Typeset by MPS Limited, Chennai, India



List of contributors

Mick Arnold Cornelis Adriaansens, Te Kura Pūtaiao Koiora—School of Biological Sciences, Te Whare Wnanga O Waitaha—University of Canterbury, Ōtautahi, Christchurch, New Zealand

Emma Aronson, Department of Microbiology and Plant Pathology, University of California, Riverside, Riverside, CA, United States

Mainak Bardhan
The John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, FL, United States
The Dr. John T. Macdonald Foundation, Department of Human Genetics, University of Miami Miller School of Medicine, Miami, FL, United States

Dominick A. Centurioni, New York State Department of Health, Biodefense Laboratory, Division of Infectious Diseases, Wadsworth Center, Albany, NY, United States

Chaitali Dagli, Department of Epidemiology, School of Public Health, University of Alabama at Birmingham, Birmingham, AL, United States

Jessica L. Davies, Oxford Nanopore Technologies plc, Gosling Building, Edmund Halley Road, Oxford Science Park, Oxford, United Kingdom

Alexander J. Diaz, New York State Department of Health, Biodefense Laboratory, Division of Infectious Diseases, Wadsworth Center, Albany, NY, United States

Brenan Durainayagam, Metabolon, London, United Kingdom

Julia Eales, Oxford Nanopore Technologies plc, Gosling Building, Edmund Halley Road, Oxford Science Park, Oxford, United Kingdom

Page xii **Christina T. Egan**, New York State Department of Health, Biodefense Laboratory, Division of Infectious Diseases, Wadsworth Center, Albany, NY, United States

Annie M. Evans, Metabolon, Morrisville, NC, United States

Jialin Gao, Division of Biostatistics, Department of Population Health Sciences, Weill Cornell Medicine, Cornell University, New York, NY, United States

Kelli D. Goodman, Metabolon, Morrisville, NC, United States

Caz Harrold, Oxford Nanopore Technologies plc, Gosling Building, Edmund Halley Road, Oxford Science Park, Oxford, United Kingdom

Shaomei He, Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, United States

Craig William Herbold, Te Kura Pūtaiao Koiora—School of Biological Sciences, Te Whare Wnanga O Waitaha—University of Canterbury, Ōtautahi, Christchurch, New Zealand

Kerry L. Ivey
Division of Aging, Department of Medicine, Brigham and Women's Hospital, Boston, MA, United States
Department of Medicine, Harvard Medical School, Boston, MA, United States

Jacques Izard, Department of Internal Medicine, University of Nebraska Medical Center, Omaha, NE, United States

Adam D. Kennedy, Metabolon, Morrisville, NC, United States

Ziyu Liu, Division of Biostatistics, Department of Population Health Sciences, Weill Cornell Medicine, Cornell University, New York, NY, United

States

Himel Mallick
Division of Biostatistics, Department of Population Health Sciences, Weill Cornell Medicine, Cornell University, New York, NY, United States
Department of Statistics and Data Science, Cornell University, Ithaca, NY, United States
Division of Gastroenterology and Hepatology, Weill Cornell Medicine, Cornell University, New York, NY, United States

Rohan Mishra, BioRankings, St. Louis, MO, United States

Michael J. Perry, New York State Department of Health, Biodefense Laboratory, Division of Infectious Diseases, Wadsworth Center, Albany, NY, United States

Aaron Pomerantz, Oxford Nanopore Technologies plc, Gosling Building, Edmund Halley Road, Oxford Science Park, Oxford, United Kingdom

Daniel M. Portik, Pacific Biosciences of California, Inc., Menlo Park, CA, United States

Maria C. Rivera, School of Life Sciences and Sustainability, Virginia Commonwealth University, Richmond, VA, United States

William D. Shannon
BioRankings, St. Louis, MO, United States
Washington University School of Medicine, St. Louis, MO, United States

Jeremy E. Wilkinson, Pacific Biosciences of California, Inc., Menlo Park, CA, United States

Naupaka Zimmerman, Department of Biology, University of San Francisco, San Francisco, CA, United States

Preface

The second edition of *Metagenomics for Microbiology* provides an opportunity to reflect on the growth of the metagenomics field, from new applications to technological advances. It also articulates current challenges and the need for technological developments to facilitate new solutions and applications. In each chapter of this edition, the authors discuss their areas of expertise and the resources available for implementing omics technology across various fields of biological science.

The application of omics technologies in the characterization of microbial communities, including metagenomics, transcriptomics, metabolomics, and proteomics, continues to expand across diverse scientific domains. These methodologies offer unparalleled insights into biological systems, enabling high-resolution exploration of microbial communities by providing insights into their diversity and composition, gene expression patterns, metabolic pathways, and functional capabilities. The use of omics technologies is so widespread in microbiology that it would take multiple books to list all its applications. As editors, our primary goal was to compile a volume highlighting the essential topics that would provide scientists new to the field with the background information to facilitate the design, implementation, and analysis of microbiome studies employing omics technologies. Another goal was to provide experienced scientists with easy access to bibliographical information.

This book contains a series of chapters from leading experts describing best practices for collecting and analyzing omics data, as well as highlighting the

field's promises. **Chapter 1** provides an overview of the omics pipeline from end to end and some solutions for hurdles that might be encountered. **Chapter 2** offers insights into tools, trends, and trajectories from both the past and present, as well as how future technological advances could reduce present-day challenges.

The following chapters highlight the expansion of the field. In **Chapter 3**, the authors discuss their experience in the application of

metagenomics to microbial biothreat and emerging threats. Meanwhile, **Chapter 4** highlights the central role of symbioses in the evolution of life by examining how host–microbiota associations have evolved and diversified along the endosymbiont–microbiome continuum. It emphasizes the diversity and distribution of these symbiotic relationships throughout the tree of life. **Chapter 5** is an introduction to the essential statistical concepts required for proper metagenomics experimental design and emphasizes the importance of an appropriate sample size that allows statistical analyses to draw meaningful conclusions. **Chapter 6** gives the basis to maximize transcriptomics sequencing investment by removing the ribosomal RNA. The following two chapters discuss the most recent advances in applying next-generation long-read sequencing technologies to metagenomics and transcriptomics. **Chapter 7** describes the approaches developed by Oxford Nanopore Technologies, while **Chapter 8** discusses the PacBio HiFi sequencing applications. **Chapter 9** introduces the concept of metagenome-assembled genomes, commonly called MAGs, which provide a solution for discovering genomes present in metagenomic samples when cultivation or isolation of organisms is not feasible. **Chapter 10** challenges the reader to think about how the data is distributed in function of the exact nature of the components of the omics datasets. **Chapter 11** presents developments in the characterization of microbiomes by the study of small molecules via metabolomics. The availability of complex multiomics datasets (metagenomics, transcriptomics, metabolomics, and proteomics) requires specialized analytical tools. **Chapter 12** discusses the application of

supervised machine learning, unsupervised approaches, and multimodal AI (artificial intelligence) techniques for integrating multiomics datasets.

The advances in the microbiome field are the result of interdisciplinary collaborations and investments by governments, the private sector, and nongovernmental organizations (NGOs). The complexity of microbiome systems requires renewed efforts by all stakeholders to overcome the present obstacles. By addressing these challenges, omics technologies can continue to revolutionize fields such as environmental microbiology, personalized healthcare, agronomy, agriculture, aquaculture, and synthetic biology; this will help us develop innovative solutions to global scientific and societal problems.

Page xvii This book's integrative approach will appeal to researchers, educators, students, and citizen scientists interested in exploring microbial metagenomics. For microbiologists generating microbiome omics data, this book offers an introduction and guidance on the computational and statistical aspects of data analysis. For statisticians and computational scientists considering work in this field, this book provides foundational knowledge of the data types they aim to model. More broadly, this book serves as a starting point for the scientific community to delve deeper into this transformative and captivating area of study.

We thank the authors for taking the time to develop their contributing chapters. We acknowledge the public and private funding organizations whose support has made these technological and conceptual advancements possible. Additionally, we thank the pioneering researchers and consortia whose efforts laid the groundwork for these innovations to thrive. Finally, we are grateful to Elsevier for entrusting us once again as editors for the second edition of *Metagenomics for Microbiology*.

Jacques Izard

Maria C. Rivera

CHAPTER 1

Toward success in omics studies

Jacques Izard, Department of Internal Medicine, University of Nebraska Medical Center, Omaha, NE, United States

Abstract

Metagenomics, under all its forms, is a powerful practical, technical, and conceptual approach to observing and inferring biological phenomena via high-density nucleic acid sequencing and analysis. Sequencing, by itself, does not solve any question; it must be embedded in a succession of steps or pipelines, from design, sample collection and preparation, sequencing, and analysis. As with all technical pipelines it comes with tremendous advances, new opportunities, limitations, and biases. This overview focuses on some parameters that need to be considered in the planning and execution of omics experiments. Whether specialist or novice in the field, traditional researcher or citizen scientist, there are a few things to keep in mind when designing and implementing a project.

Keywords

Metagenomics; metatranscriptomics; shotgun sequencing; WGS; DNA-Seq; RNA-Seq; 16 S rRNA gene; omics; study design; bioinformatics

Metagenomics and why we consider it a pipeline

The general definition of metagenomics is the cultivation-independent analysis of genetic information from collective genomes and genetic material within a given environment based on its sampling (Berg et al., 2020; Handelsman et al., 1998; Izard, 2015). The approach is rooted in nucleic acid

sequencing technologies. DNA, RNA, or both are the target of the study. The diversity of hypotheses, questions, and goals to be addressed is endless.

Metagenomics is tightly associated with the availability of next-generation sequencing and its associated depth of nucleotide sequencing per sample. Much more data are acquired per sample at a lower price over the evolution of the technology. The additional information available for each sample and thus to the overall study changed the traditional paradigms of "What is the most abundant?" or "What is the most significant?" to "What is present and potentially significant that might influence the situation and outcome?"

This characterization and quantitation of pools of nucleic acids and other biological molecules can be characterized in broad strokes under the following questions:

- Who is there? A classic example is the use of 16S rRNA gene sequencing to identify bacterial taxa in a sample. Whole genome-shotgun sequencing can provide excellent taxonomic resolution, for all domains of life, with the right tools.
- What can they potentially do? This is the world of genes and their function. It is the intersection between gene discovery, identification by sequence similarities, and function assignment via comparison with conserved sequences of known functions. A classic example is Page 2 the use of this approach for the evaluation of the spread of antibiotic resistance. It is traditionally the domain of DNA sequencing but also bleeds into the next question.
- What can they really do? The transcriptome is the closest step to functional action at the level of nucleic acids. This classical definition is challenged by the myriad of small RNAs that can be now detected, but the overall concept prevails. Transcriptomics is the sequencing of RNA molecules expressed by organism(s). Like functional genomics, similarities drive identifications. As not all genes are transcribed, the transcriptome is the closest nucleic test to what molecules are being

produced by the cells. As not all genes are functional, not all transcripts lead to a protein or peptide.

- What are they doing? While conceptually the answers can be derived from DNA and RNA sequencing, this is the domain of molecule identification by metabolomics, lipidomics, proteomics, etc.
- What are the potential molecular interactions? All molecules in the cell interact transiently or permanently. Nucleic acids, small molecules, and proteins form complexes and molecular machinery necessary for life functions.
- What is the context of the biological activity? To each previous question, one should add the contextual data (type of sampling site, environment, etc.), the biogeography (human descending colon biopsy, rhizome of the orchid *Orchis provincialis* at the following GPS coordinates, etc.), timeline (cross-sectional or longitudinal study, date and time, etc.), metadata (information gathered about the study site, phenotypes, etc.), and metadata of the sample for troubleshooting failure or validating signal (sample preservation characteristics, extraction kit lot number, etc.).

None of the questions above can be answered without a systematic approach ensuring rigor, reproducibility, and quality at each step (**Fig. 1.1**). Each step has its own requirements for standardization to maximize output quality while decreasing variability. Most of the steps are sequential, where the output of one step is the primary material for the next one. The biological pipeline requires appropriate collection and storage of the samples depending on the study requirement, the environment, the sample extraction methodology, and the sequencing platform to maximize data output and effectively answer the hypothesis. For instance, when studying the sponge microbiome in both Antarctica and a laboratory tank, the standard operating procedure Page 3 will differ drastically in several aspects, despite using the same kits. If we take the example of metadata collection that is contiguous to

sample collection, a data curation step is required to assert quality and prepare the data for statistical and biocomputing needs (e.g., format of the dataset to match software data input requirements and coding of the data to satisfy categorical or continuous formatting). Defining your processes as a set of pipelines will streamline discussions and provide further support for reproducibility. Page 4

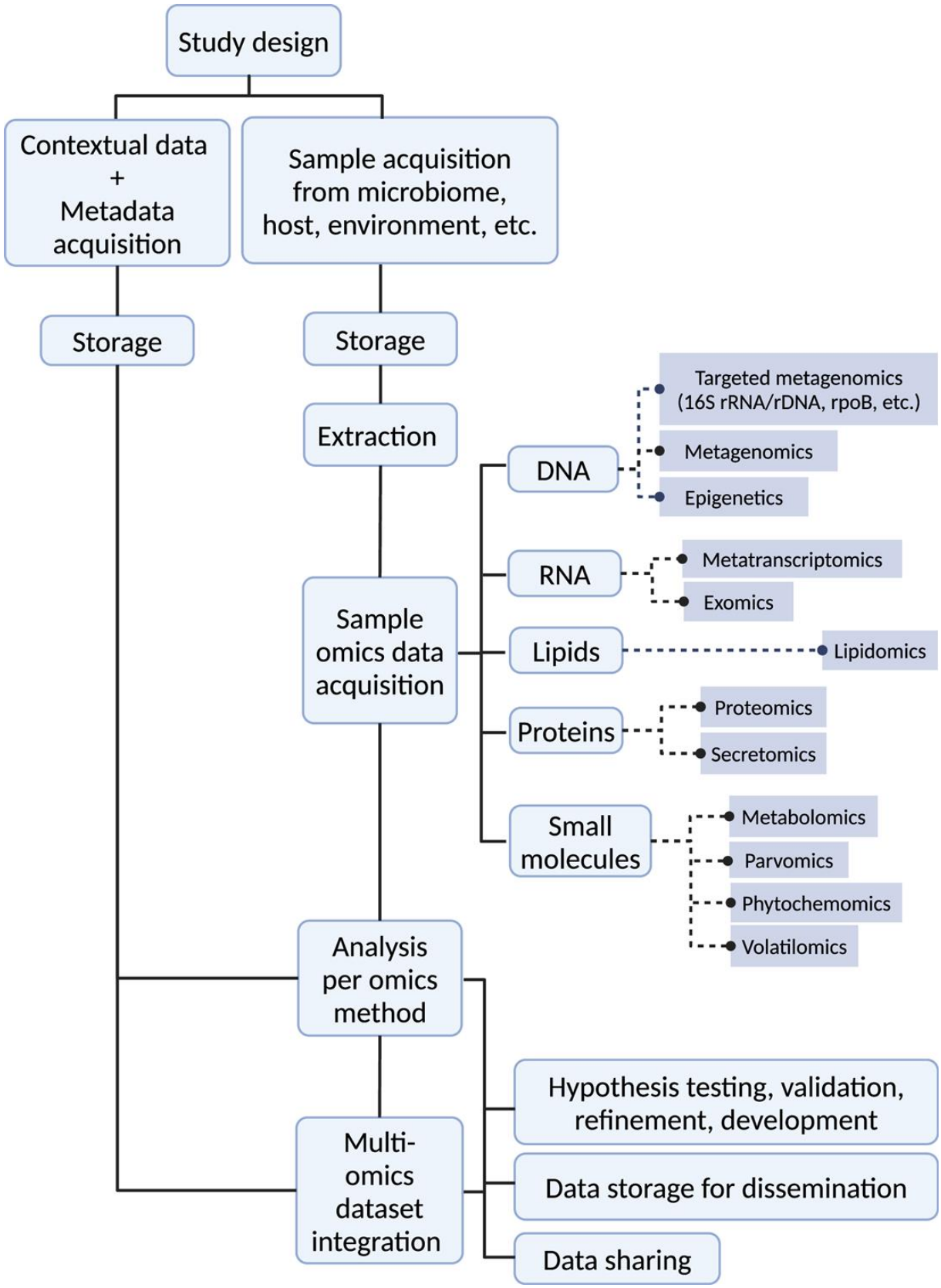


FIGURE 1.1 Holistic view of an omics study from study design to data sharing. Created with BioRender.com.

Follow for extended description

Study design: aligning your steps to improve overall quality

The goal is to emphasize some aspects of study design that need careful attention due to the characteristics of the methodologies used in metagenomic studies. It begins by identifying the primary objective of the metagenomics project. What is the main scientific question you are trying to answer? More than one hypothesis can be tested depending on the scope of the experiment and the amount of associated data, or metadata, that you collect and use for your subsequent analyses.

Pragmatically a well-designed metagenomics pipeline will enhance feasibility, avoid bottlenecks, optimize monetary and human capital utilization, promote reproducibility, and improve your chance of higher-impact data dissemination. However there is no secret about it. The design must be done prior to the start of the experiment. Adjustments will be necessary but having an operating procedure provides leverage within the team and for external collaborators and funding partners.

While we cannot list all the parameters involved in study design, below is a selection of the obvious and not-that-obvious components to consider. Host-associated microbiomes and environmental samples often appear to have distinct requirements; they share fundamentally similar challenges.

Statistical analysis plan: an early step, not an after-thought

Planning for statistical analysis should be an integral part of the study design. Many experimental designs can be performed for metagenomics projects. There is no single path to a successful strategy. While using metagenomic or metatranscriptomic approaches, it is essential to refer to the specific needs of each experiment.

The statistical analysis plan should take into account the characteristics of the experiment (in human studies, this would be the inclusion and exclusions criteria), the rate of sample acquisition (this would include the rate of human subject recruitment which will determine if you are working with one or more batch of datasets), the descriptive objectives, testable hypotheses, the statistical methods that might be stand-alone or embedded in bioinformatics tools or pipelines, etc. One of the direct advantages of planning ahead is that when you have the data in hand, you'll have a strategy in place to start the analysis. This is critical as next-generation sequencing provides a tremendous amount Page 5 of data and you want to remain focused on your primary objective(s). After the completion of your primary objective(s) exploratory analyses and investigation or formulation of additional hypotheses are always possibilities.

The statistical plan will also include how many samples you should collect (sample size). The most basic question about the research plan should be: "Are enough samples being collected from each site or from enough subjects to make meaningful conclusions?" To properly assess the degree of similarity or dissimilarity between bacterial communities a measurable difference, or effect size, is necessary. The effect size refers to a quantitative measure of the magnitude of a difference, association, or impact between groups, conditions, or variables. It provides an indication of how substantial the observed difference is, independent of the sample size. In general the smaller the effect size and the greater the variability within a group of samples, the larger the number of samples required to achieve adequate statistical power (Kelly et al., 2015; Kers & Saccenti, 2022; Knight et al., 2018; Rosa et al., 2015).

You might not find a metagenomic dataset to help or guide you in the experimental design phase. Instead previous studies using other molecular techniques or culture-based methods might be used as a reference to design and plan your own experimental design. If you were looking at the same question with a more traditional method, you should have enough samples, at the proper sequencing depth, to detect differences if they are present.

To that end there are many different types of study designs with important considerations about the number of samples you should collect. The cross-sectional study design is an observational study design that analyzes data and samples from a population, or a representative subset, at a specific point in time.

Unlike longitudinal studies that analyze samples over a predetermined time frame cross-sectional studies provide a snapshot of the variables of interest within the study population at one particular moment. This design is commonly used to assess the prevalence of conditions, identify associations between variables, and generate hypotheses for further research. However, due to its temporal limitations, it cannot establish causal relationships. Longitudinal studies present their own challenges and can be analyzed at each time point independently, along the timeline and across body sites (Björk et al., 2024; Bogart et al., 2019; Gerber et al., 2012; Zhou et al., 2019). Page 6

Ethical and legal issues

For human studies the Council for International Organizations of Medical Sciences (CIOMS), in collaboration with the World Health Organization (WHO), publishes guidelines for health-related research involving humans, including guidelines on ethical issues in medical genetics and genetic services (World Health Organization, 2016). In the United States the National Human Genome Research Institute (NHGRI), part of the National Institute of Health, offers resources and information on ethical issues in genetic research on its website.

Regardless of your study design, from observational to clinical trials, a review by Institutional Review Boards (IRBs in the United States; name may vary locally) is required before research can begin. The consent of a knowledgeable individual is key. Participants' willingness to share their time and personal information enhances your study design and further facilitates your ability to contextually analyze the data. This information usually

includes private details. Therefore, the needs related to the protection of the identifiable health information of human subjects are critical (e.g., HIPAA Privacy Rule in the United States; name may vary locally). Part of the samples collected may also enable the researchers to put the genetic information of the participants in the context of the research question. The guidelines for genetic research vary per country. Work with your local Institutional Review Board to maximize equity and justice, cultural sensitivity, and transparency. Genetic counseling might need to be included. The CIOMS, in collaboration with the WHO, publishes international ethical guidelines for health-related research involving humans, including guidelines on ethical issues in medical genetics and genetic services (World Health Organization, 2016). In the United States the NHGRI, part of the National Institute of Health, offers resources and information on ethical issues in genetic research on its website.

For animal studies animal welfare is a key component, and protocol review by the appropriate institutional regulatory body is key (e.g., IACUC in the United States; name may vary locally). The principle of the 3 Rs of animal research does apply to metagenomic studies by replacing the use of animals whenever possible, reducing the number of animals needed to a minimum, and refining the tests to cause the animals the least amount of distress (Guide for the Care & Use of Laboratory Animals, 2011; Russell & Burch, 1959; Tannenbaum & Bennett, 2015). However not all organisms are covered by regulatory bodies. Page 7

For environmental samples, the right to access land or bodies of water may require a sampling permit. Working in areas with endangered species provides its own requirements. In any case ethical precepts should not be limited to when regulations exist. A set of guiding principles based on the idea that we are all in the same ecosphere called Earth has been published (Lange et al., 2022).

The so-called "contamination by the host and/or environment"

The presence of host or environmental nucleic acids, in most instances, is not a contamination; rather, it is part of the sample. It might cost you more to get the proper sequence coverage, but microbiomes are inherently involved in their surroundings. This is to be distinguished from true contamination which are unwanted impurities unrelated to the studied system. Contamination will be detected as an integral component of the sample, due to the depth of the data being acquired. Sophisticated computational pipelines exist for the removal of unwanted reads. This removal step is essential to generate meaningful conclusions based on the microbiome. In the case of studies involving human participants this step also protects the privacy and confidentiality of the donor.

The percentage of human reads in samples from 16 different body sites was investigated during the Human Microbiome Project (Human Microbiome Project Consortium, 2012; Izard, 2015). The body sites could be divided into five categories: skin, nasal, vaginal, oral, and stool. The nasal and vaginal sites, as well as saliva samples, had the most consistent high-level human reads. The tongue dorsum, oral hard palate, and stool samples consistently had the lowest levels of human reads, resulting in a greater number of microbiome-related reads to analyze for the same depth of sequencing. Some methods to decrease host nucleic acid content from samples exist (Wu-Woods et al., 2023).

The same concepts apply when working with nonhuman systems. In addition to the biogeography of a biological system it should be noted that the genome of a host might be similar in size to a human or vary widely. For reference the human genome is about 3.2 billion base pairs (Schneider et al., 2017; Venter et al., 2001). *Rattus norvegicus*, or the common brown rat, is one of the most well-established experimental model organisms, and its genome sequence is 2.4 billion base pairs (Howe et al., 2021; Modlinska & Pisula, 2020). The same size of genome was observed for corn

(Haberer et al., 2005). Protists and fungi are much smaller with genomes composed of a few million base pairs. In contrast the Desert Locust (*Schistocerca gregaria*) genome is 8.8 billion base pairs (Verlinden et al., 2020). The knowledge of your biological system of interest will be critical to optimize the study design.

Sample acquisition and storage: key steps to get the best data

Although the technology of the sequencing platforms has evolved, their focus remains on sequencing the nucleic acids, whether DNA or RNA. The source of the microbiome sample varies greatly, ranging from environmental samples to plants, insects, fish, animals, and humans. Published data on environmental samples are diverse, including soil, hot springs, seawater, air, and surfaces in homes and hospitals. For plants the above- and below-ground microbiomes have been well studied. For insects animals, including humans, investigations have been conducted on multiple body sites.

In the sample acquisition and following steps, many factors will contribute to the decision-making process. These include the hypothesis involved, the goals of the project, the available facilities and personnel, and the available funds. Planning these steps prior to the initiation of the study is invaluable in ensuring an efficient and successful study.

The proper sampling protocol is essential to a successful metagenomics study, since the accurate identification of many organisms depends on the collection and handling of the sample. Defining the geographical location or the specific body site, surface, depth, volume, or quantity to be collected is necessary for sampling standardization. If possible, keeping the samples concentrated, and processing them for immediate storage are often key. Protecting the samples against the elements (wind, sun, etc.) sounds like obvious advice. Desiccation of samples is a common problem when working with small samples and does not require exposure to the elements or a drastic event.

The nature of the sample often dictates what is the best protocol to avoid sample desiccation, denaturation, lysis, degradation, etc. As immediate extraction on-site or access to an −80°C freezer is not always an option, alternatives must be developed to preserve the sample, its integrity, and its value for the question(s) at hand. From self-made kits to standardized commercialized kits the scientist has many options. Page 9 Whether in the middle of the Amazon or in the bathroom of a participant, kits with fixative preserve the sample as a snapshot at the time of collection (Franzosa et al., 2014; Guan et al., 2021; Isokääntä et al., 2024; Zouiouich et al., 2023). However be aware that some fixatives are not compatible with some other techniques like metabolomics (Guan et al., 2021; Isokääntä et al., 2024). These kits are valuable if you collect a few samples or more than 20,000 samples, as they ensure standardization and reproducibility (Everett et al., 2021). Additionally studies now frequently collect biological samples across time and/or geography; thus a method compatible with different lengths of transport is critical.

It is recommended to examine each step and consider potential inadvertent contamination by exogenous nucleic acid material, or potential inhibitors for downstream applications. This is particularly applicable for tools and equipment that are reused, where proper cleaning and sterilization procedures are essential. The following guidelines are simple ways to increase the quality of sample preparation. Not talking over a biological sample or wearing a facemask would eliminate contamination by the technician's breath. While protecting the sample using gloves, do not forget that a simple touch of the skin or a surface would contaminate the glove which in turn might contaminate the sample itself. Natural DNAses and RNAses may potentially damage the sample. It is most often about applying common sense in the context of the depth of the data to be gathered. In other words, if you want to know the microbiome of the banana peel on the plant but drop the banana in the field, you are also going to learn about the microbiome of that square of Earth and that of the fruit surface.

Analyzing true and technical replicates of a sample and assessing whether observed differences are statistically significant is a good practice. True replicates, which occur when the same site is sampled more than once, are rarely done in metagenomic studies because the sensitivity of the technique may easily show differences when sampling a site multiple times due to the biological organization of the site (Zeeuwen et al., 2012). In contrast technical replicates, where the sample is split for processing, are easy to perform for reassurance that sample handling is not the source of the observed variation.

Storage and sampling are tightly linked issues. It is not always possible to have a freezer or an expert on location when the sample is collected. Solutions for these problems affecting the downstream steps need to be identified before starting the study. The nature of the type of

sample is too diverse to enter all the details, but one key question will drive the process: “How many samples do I really need per application?” The associated questions would be: “Do I need DNA, RNA, proteins, lipids, small molecules, etc., from the same sample?” and “Will the sample be used for more than one application, preparation, or extraction?” These along with any other questions related to the present or future study applications should be considered early on to acquire the best sample to answer the question(s) you are investigating.

Maintaining consistency in the sample collection and storage protocols will preserve both the quality of the sample and limit the batch effect during the analysis, enhancing the signal and minimizing data noise. A batch effect is the nonbiological differences introduced into the data due to variations in experimental conditions, rather than true biological variation. This includes variations in sample storage conditions (in particular in the absence of a fixative), in laboratory equipment used (e.g., different centrifugation speeds between samples), reagents (e.g., not using the same kit across the sample extraction), or in data handling (e.g., not using the same version of the software for different group of samples), etc. Sometimes, you cannot avoid it because a sampling kit has been discontinued or the centrifuge broke. Take

note of these variances and keep them in mind during the data analysis. Batch effects can significantly impact the interpretation of microbiome data, leading to false signals or masking true biological differences.

Sampling cost is often neglected. You might have multiple steps in your process to reach the final storage space, and there is no issue with that. However it should also be stressed that whether you work on a large human subject cohort or a large field collection, the costs of personnel, sampling equipment (single use when possible), and transient as well as permanent storage add up quickly. Regardless it is key to optimize your process to be the most consistent for each sample or per batch of samples. With the sample collected and in storage nucleotide extraction will be the next step.

Sample extraction: a step closer to high-quality datasets

The sample source in a metagenomics pipeline can be extremely diverse. DNA and/or RNA need to be extracted from the sample prior to any analysis. The type and source of the sample will determine the most appropriate extraction protocol. This step, simplified by the availability of nucleic acid extraction kits, is crucial to the success of the Page 11 analysis, as the quality of extracted DNA and/or RNA will influence all subsequent steps.

Before selecting the most appropriate extraction protocol, perform a careful review of the literature and validation of the protocols for your specific sample. The choice of protocol will depend on the DNA or RNA yield, shearing, removal of contaminants (which could be inhibitory to subsequent steps), and extraction bias. The availability of kits that allow the purification of more than one product per extraction, that is, nucleic acid or macromolecule, provides new avenues in omics studies.

Purification kits do not purify all the molecule types that are on the label. An example of this issue is the case of microRNA purification. These functional RNAs are often washed from the column during purification. Their importance is undeniable, however, rarely the focus of the experiments

(Boivin et al., 2019; Francavilla et al., 2023; Tarallo et al., 2022). Specialized kits exist, check which one match your needs. Along the same lines, RNA Poly-A tails exist in the microbiome. Be aware of kits selecting by poly-A or removing the poly-A (Altmäe et al., 2020).

Sequencing: an equation between hypothesis, sequencing depth, and financial matters

The cost associated with the production of a megabase of raw reads has been decreasing steadily over the years (Wetterstrand, n.d). Starting around $10,000 at the turn of the century using Sanger-based DNA sequencing, it dropped to $100 by 2008 with the first implementations of next-generation sequencing. By the time of the Human Microbiome Project in 2011, the price had decreased to $0.1, and by 2021, it was about $0.01. Don't be fooled by the apparent low cost. This pricing does not include library construction, consumables, labor, and data processing.

Which sequencing platform to use? Due to the varied nature of scientific studies, there is no single approach that is recommended. A detailed review of the literature, discussions with colleagues and sequencing facilities, cost, availability, turnaround time, and scope of the project will be part of the decision-making process. Let's not forget that the hypothesis and the goal should be the true drivers of technology selection.

How much sequencing depth is needed? Determining the number of reads required is a tradeoff between the minimal number of reads needed to allow an informative and statistically significant analysis and Page 12 the available budget. This choice is driven by both the platforms and your experimental needs such as the previous knowledge of the relative abundance of your organism(s) or metabolic pathway(s) of interest. If your metagenome or metatranscriptome is of relatively low complexity, you can use available genome sequences to evaluate the coverage needed (Haas et al., 2012). For a metatranscriptome you'll have to adapt the sequencing coverage if your focus is the most abundant transcripts or the rare transcripts. It has been shown

that millions of 16S rRNA reads do not appreciably increase the extracted information and that a cost-efficient read number is sufficient to discriminate adjacent sites (Caporaso et al., 2011; Rosa et al., 2012).

When working with human samples, all efforts should be made to secure information and confidentiality of the genetic material and associated data over time. This includes the storage of the information, data deposition, and data sharing, even when the samples are de-identified.

Metadata: a context for the analysis and the door to our overall understanding

Alongside sample acquisition the acquisition of metadata is crucial for statistical analysis and study interpretation. The quality and extent of the metadata are critical to any metagenomic study. Metadata will enhance your analysis beyond the most obvious evidence. It provides context to the experiments and allows for meaningful comparisons between studies, while deepening our understanding of the dataset. With a greater depth of information a broader knowledge of the "environmental factors" is needed. While not the focus of an experiment, seemingly extraneous data may become important. For example information on the source of carbon for microbial metabolism might be later identified as a confounding variable in an experiment. It can be as simple as the source of sugar intake for a subject or the nature of the pollutant for a soil sample.

The information about the sample location or its relative location to other samples can be included in the analyses. The concept of biogeography goes beyond the description of environmental features that influence the spatial distribution of microorganisms. It aims to understand the metabolic processes within the microbes' own niche and their relationships with other biological niches. The niche might be the different sites in the oral cavity, along the digestive tube, or in the skin (Grice et al., 2008; Segata et al., 2012; Zhang et al., 2014). Large-scale data visualization and analysis tools have been created to help us picture these aspects (Gonzalez et al., 2012).

As we are discovering the microbiome as an interdependent organ of any biological system (McFall-Ngai et al., 2013), we need to redefine what are the best-associated data to collect along with the sample (Kumar et al., 2024). Defining or re-defining the phenotype of interest might have a crucial importance. Because the phenotype is the result of the interaction between the genotype and the organism's environment in all its complexity, including the microbiome, we are required to renew our attention to the granularity of the defined phenotype. From the macro to the molecular scale new considerations that were previously neglected due to the lack of significance might be at play when scrutinized with a different sliver or window of observation.

Data analysis: finally getting the information you have been waiting for

Next-generation sequencing moved us from the kilo- and mega-bytes size files to the mega- and terabytes size file world. Although this might not be of great importance when you are performing a single metagenomics experiment, it can quickly become an issue in large-scale studies. While some steps can still be done on a laptop, others require a computer cluster or cloud computing. This can be a significant added cost. Be sure that the data and related information are secured during transfer and in the cloud.

Once again your application, samples, and questions will drive the computing pipelines that you or your collaborators will be using. Many computing pipelines and packages have been created, deposited, and made available in public repositories. Check that they are updated, or that the databases they are using are updated. While some packages and biocomputing resources endure the passage of time (Beghini et al., 2021; Blanco-Míguez et al., 2023; Chen et al., 2023; Gautam et al., 2023; Keefe et al., 2023; Odom et al., 2023; Schloss, 2020; Segata et al., 2011; Weber et al., 2018), some do not.

A simple search in the bibliographic database PubMed from the National Library of Medicine or Google Scholar leads to numerous articles and reviews regarding all the steps required from raw data to fully analyzed datasets. Contextual analysis is key, thus be sure to understand the parameters associated with your study. Page 14

Be aware that the metagenomics, transcriptomics, and targeted metagenomics (16S rRNA gene or other genes) data are compositional (Gloor et al., 2017). In other words, in these omics, the total read count observed is of a fixed size, representing a random sample of the relative abundance of the molecules in the sample. Unfortunately, in almost all cases, the count cannot be related to the absolute number of molecules in the original sample. This must be taken into account in the analysis.

A key component of the computing resources is the use of peripheral databases to decipher taxonomy, phylogeny, and function. Some of the databases associated with phylogeny, function, or repository of nucleic or protein sequences include the Human Oral Microbiome (HOMD) (Chen et al., 2010), Greengenes2 (McDonald et al., 2024), OceanDNA metagenome-assembled genome (MAG) (Nishimura & Yoshizawa, 2022), Clusters of Orthologous Genes database (Galperin et al., 2018), KEGG for taxonomy-based analysis of pathways and genomes (Kanehisa et al., 2023), carbohydrate-active enzyme database (CAZy) (Drula et al., 2022), GenBank (Sayers et al., 2024) as well as other international public depositories, and curated databases.

Some computing pipelines offer the choice of database to be used as reference, while others do not. Be aware that none of the databases are identical. With the addition of MAG analyses can differ from the traditional results by expanding our horizon of what is known. MAGs are genomes assembled from metagenomics reads. After assembly the contigs are binned based on compositional profiles allowing the formation of larger contigs or scaffolds. This process does not relay on the knowledge from the sequence of a purified/isolated/sequenced bacteria or another member of the

microbiome. MAGs have been asserted to be a close representation of an actual individual genome of an already existing isolate or represent a novel isolate.

Details frequently overlooked (snippets)

Defining a healthy control is more challenging than it appears. For example the Human Microbiome Project checked for the absence of inflammation at 18 body sites. Each participant was screened by multiple specialists (Human Microbiome Project Consortium, 2012a; Human Microbiome Project Consortium, 2012b). This is out of the scope of most investigators. Another instance that exemplifies this challenge is the rise of obesity (Centers for Disease Control & Prevention. Adult Obesity Prevalence Maps, 2022; NCD Risk Factor Collaboration (NCD-RisC), 2017). Page 15 Are individuals l parameters, or because they represent a so large part of the population, do they represent the expected control population? This is a contextual decision to be made.

No microbial consensus systematics has been agreed upon. An International Code of Nomenclature of Bacteria (Bacteriological Code) is maintained for prokaryotes. It contains general considerations, principles, rules, and recommendations that govern the way in which the names of prokaryotes are to be used (Arahal et al., 2023; Oren et al., 2021; Oren et al., 2023). A similar approach was developed for the fungome (Hyde, 2023). However different databases might use different systematics. It is a problem for biocomputing that is not new (Stackebrandt, 2007). Choose a database and complement it as needed. The easiest way to check for the spelling and/or the validity of a name, including its historical evolution, is by checking the List of Prokaryotic names with Standing in Nomenclature (Meier-Kolthoff et al., 2022; Parte et al., 2020). Be aware that the name of the phyla has changed since 2021 (Oren & Garrity, 2021).

No standard definition of a strain exists. A strain is traditionally defined based on isolates. The differentiation can be derived from genetic (plasmid,

gene, genome, etc.), metabolic (based on panel or single metabolic difference, production of a valuable metabolite, etc.), phenotypic (shape, colony morphology, biochemical properties, toxin production, antibiotic resistance, pathogenicity, etc.), or antigenic (via serotyping, agglutination, etc.) characteristics, as well as ecological niches and host range. An effort to standardize the definition of a strain is being made using biocomputing with variable level of success (Ghazi et al., 2022; Heinken et al., 2023; Mei et al., 2024).

Genome size is variable and contains a strong phylogenetic signal, but it is not limited to inter-kingdom differences. The diversification observed is still influenced by shared evolutionary histories (Martinez-Gutierrez & Aylward, 2022). A factor of over 10 exists between the smallest and the largest genome sizes in bacteria and archaea (Martinez-Gutierrez & Aylward, 2022). Evaluating genome coverage, that is, how many times a segment of a genome is fully sequenced in the dataset, implies a good understanding of microbial representation within a sample.

Genome copies are a moving target associated with the dynamics of the population and the growth phase. Not all bacteria conform to the patterns exemplified by *Escherichia coli*, which genome is distributed in Page 16 one chromosome and has only one copy per cell. Rather we must contend with the fact that the genome copy number can change in the different phases of growth and that more than one ploidy can be observed in a population (Caro et al., 2007; Müller & Babel, 2003). The biological significance of polyploidy will depend on the system studied (Griese et al., 2011; Komaki & Ishikawa, 2000; Pecoraro et al., 2011). Additionally the cells can replicate asynchronously, displaying a heterogeneous DNA content (Müller & Babel, 2003; Müller & Nebe-von-Caron, 2010). The role of polyploidy and replication in influencing the phenotypic features and, subsequently, the data from each "omics" platform are still an open question.

Intraspecies microbial diversity is a reality. Each microbe is subject to a variety of evolutionary processes. These include mutation, selection, drift,

and selective pressure from xenobiotics (Cullen et al., 2020) In complex environments metabolic potential, among others, can be changing across geography or biogeography. Modern genomics is extremely powerful but does not replace a good understanding of microbial biology.

The number of 16S rDNA is variable per organism. The number of 16S rRNA genes can vary from 1 to approximately 15. Within species variation also exists. Two strains of *Bacillus subtilis* from the Bacillota phylum (previously known as Firmicutes) differ by two copies, while their genome differs in size by only 4% (del Castillo & Izard, 2015). While there is a weak correlation between genome size and the number of 16S rRNA genes, there is a stronger correlation with the number of ribosomes that a cell can produce simultaneously, and this reflects the ecological strategy in microorganisms' ability to adapt quickly to fluctuating environmental conditions (Klappenbach et al., 2000; Martinez-Gutierrez & Aylward, 2022; Niederdorfer et al., 2017).

Reference strains exist in larger diversity than most people expect. If interested in virology, mycology, bacteriology, or protists, deposited strains exist. American Type Culture Collection (ATCC), German Collection of Microorganisms and Cell Cultures (DSMZ), National Collection of Industrial, Food and Marine Bacteria (NCIMB), and Japan Collection of Microorganisms are among the best-known depositories. Do not hesitate to look at other options to meet your specific needs.

Standardized microbial communities can be purchased and included in your experiments. They are available from sources like ATCC, Zymo Research Corp., and the National Institute of Standards and Page 17 Technology. These standardized microbiomes can be used as validation, plate-to-plate controls, quality control, etc.

Data publication and dissemination

It is expected that any metagenomics project will develop a plan for sharing the data collected with the scientific community, including sequence data and metadata. The International Nucleotide Sequence Database Collaboration

(**http://insdc.org**) hosts some of the repositories for the dissemination of nucleic acid datasets. NCBI (USA), DNA Data Bank of Japan (Japan), and European Nucleotide Archive (Europe) are the front-facing components. They provide data resources that include raw sequence reads and alignments, structured metadata describing investigated samples such as taxonomic information, experimental, and project design, assembled nucleotide sequence data with functional annotation, and sequence-derived analyses.

Due to the diversity of the ecosystem studied there is no standard format for presenting and providing metagenomics datasets. The database resources mentioned above have improved the submission pipelines to include fields that will be necessary to reprocess the data. In manuscripts the format should include the methods, tools, and parameters used in the analysis. One option is to make the information available as an online appendix to the published article.

Minimum information requirements have been proposed for both genomes and metadata to be submitted to journal and sequence repositories. The minimal information about a genome sequence, minimal information about metagenome sequence (MIMS), minimal information about marker gene sequence (MIMARKS), and minimum information about any (X) sequence (MIxS) specifications are checklists that both standardize and enhance our ability to further analyze datasets for either training or complementary analysis (Field et al., 2008; Yilmaz et al., 2011).

In the context of the interdisciplinary nature of human microbiome research spanning microbiology, biology, bioinformatics, statistics, epidemiology, and translational medicine guidelines have been published (Mirzayi et al., 2021). Strengthening The Organization and Reporting of Microbiome Studies is composed of a 17-item checklist organized into six sections that correspond to the traditional sections of a scientific publication, to be included in supplementary materials (Mirzayi et al., 2021). Page 18

There is no such thing as pressing a button and getting the dataset analyzed. Professional scientists, students, and citizen scientists encounter

the same issues. High standards should be put into service for the benefit of the biosphere and its inhabitants.

Integrating the scales

What are scales in biology and metagenomic studies? Traditionally scales in biology are described as atomic/molecular, subcellular, cellular, organismal, population, community, and ecosystem, to which we should add biomes (larger communities adapted to a climate), the Earth, and since humans left fecal material on the moon surface, we should include the solar system. Metagenomic studies subsample any of these systems.

In metagenomic studies researchers analyze genetic material recovered directly from samples, which can encompass any of the aforementioned biological scales. By examining the collective genomes and transcriptomes of microorganisms within these samples omics methods provide insights into the composition, functions, and dynamics of microbial communities across different scales.

The findings from metagenomic studies can be significantly enhanced by considering the upstream and downstream effects within the biological system under study. This holistic approach ensures a comprehensive understanding of how changes at one scale can influence processes at other levels, ultimately leading to a more integrated view of biological and ecological systems.

If one considers the soil microbiomes, they influence plant health and agricultural productivity. Integrating molecular data on microbial gene functions with organismal and ecosystem scales leads to considering how microbial interactions with plants affect crop yields and sustainability in agriculture (Bahram et al., 2018; Xiong et al., 2021).

In human studies the integration of data from the molecular scale (genes and metabolic pathways of digestive tract bacteria) to the organismal scale (human health outcomes) is a day-to-day occurrence. To that end multiomics

approaches are investigated (Beghini et al., 2021; Cheng et al., 2022; Lloyd-Price et al., 2019).

No biological system exists in complete isolation. Even organisms in a test tube behave as a community. Expand your perspective to include the various scales involved in your study and seize the opportunity to explore dynamic systems. Page 19

Examining the status quo

In science, the status quo or the existing state of affairs, often aligns with established opinions and doctrines. When new technologies challenge this status quo, resistance frequently arises, and not always in a constructive manner. It is not our role to dictate your stance on the debate concerning the advancements supported by metagenomic approaches. One clear advancement is the increased flow of data, which enhances our ability to validate hypotheses, provides opportunities for validating previously published hypotheses, and facilitates hypothesis generation.

The “old data,” published using more restricted or focused analyses, remains invaluable. As metagenomic analyses are based on compositional data, quantitative analyses have their place and greatly complement the high dimensional datasets produced by omics approaches. Moving forward, the strength of associations and mechanistic explanations are greatly enhanced by considering these approaches, providing a foundation for future discoveries.

The conceptual and practical details of each project are unique and should involve all relevant partners in the discussion and planning stages. This includes primary investigators, coinvestigators, statisticians, bioinformaticians, core facilities, providers, suppliers, and IT departments. This is a call to students, professional scientists, and citizen scientists alike: New datasets and tools are essential. Please engage in research, share your findings, and disseminate knowledge.

Acknowledgments

We would like to thank the colleagues who shared their frustrations and successes and Fan Yang for drawing (**Fig. 1.1)**. We apologize to those colleagues whose work was not included for lack of space.

References

Altmäe et al., 2020 Altmäe S, Molina NM, Sola-Leyva A. Omission of non-poly(A) viral transcripts from the tissue level atlas of the healthy human virome. *BMC Biology*. 2020;18(1):179 **https://doi.org/10.1186/s12915-020-00907-z**.

Arahal et al., 2023 Arahal DR, Bull CT, Busse HJ, et al. Guidelines for interpreting the international code of nomenclature of prokaryotes and for preparing a request for an opinion. *International Journal of Systematic and Evolutionary Microbiology*. 2023;73 **https://doi.org/10.1099/ijsem.0.005782**.

Bahram et al., 2018 Bahram M, Hildebrand F, Forslund SK, et al.... Structure and function of the global topsoil microbiome. *Nature*. 2018;560(7717):233–237 **https://doi.org/10.1038/s41586-018-0386-6**.

Beghini et al., 2021 Beghini F, McIver LJ, Blanco-Míguez A, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife*. 2021;10 **https://doi.org/10.7554/elife.65088**.

Berg et al., 2020 Berg G, Rybakova D, Fischer D, et al.... Microbiome definition re-visited: Old concepts and new challenges. *Microbiome*. 2020;8 **https://doi.org/10.1186/s40168-020-00875-0**.

Björk et al., 2024 Björk JR, Bolte LA, Thomas AM, et al.... Longitudinal gut microbiome changes in immune checkpoint

blockade-treated advanced melanoma. *Nature Medicine*. 2024;30(3):785–796 **https://doi.org/10.1038/s41591-024-02803-3**.

Blanco-Míguez et al., 2023 Blanco-Míguez A, Beghini F, Cumbo F, et al.... Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nature Biotechnology*. 2023;41(11):1633–1644 **https://doi.org/10.1038/s41587-023-01688-w**.

Bogart et al., 2019 Bogart E, Creswell R, Gerber GK. MITRE: Inferring features from microbiota time-series data linked to host status. *Genome Biology*. 2019;20 **https://doi.org/10.1186/s13059-019-1788-y**.

Boivin et al., 2019 Boivin V, Faucher-Giguère L, Scott M, Abou-Elela S. The cellular landscape of mid-size noncoding RNA. *WIREs RNA*. 2019;10(4):e1530 **https://doi.org/10.1002/wrna.1530**.

Caporaso et al., 2011 Caporaso JG, Lauber CL, Walters WA, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America*. 2011;108(1):4516–4522 **https://doi.org/10.1073/pnas.1000080107**.

Caro et al., 2007 Caro A, Gros O, Got P, De Wit R, Troussellier M. Characterization of the population of the sulfur-oxidizing symbiont of Codakia orbicularis (Bivalvia, Lucinidae) by single-cell analyses. *Applied and Environmental Microbiology*. 2007;73(7):2101–2109 **https://doi.org/10.1128/aem.01683-06**.

del Castillo and Izard, 2015 del Castillo E, Izard J. *Metagenomics for Bacteriology*. *Metagenomics for Microbiology* United States: Elsevier Inc.; 2015;113–134 **https://doi.org/10.1016/B978-0-12-410472-3.00008-7**.

Centers for Disease Control and Prevention. Adult Obesity prevalence Maps, 2022 Centers for Disease Control and Prevention. Adult Obesity prevalence Maps. (2022).

Chen et al., 2023 Chen I-MA, Chu K, Palaniappan K, et al.... The IMG/M data management and analysis system v.7: Content updates and new features. *Nucleic Acids Research*. 2023;51(D1):D723 **https://doi.org/10.1093/nar/gkac976**.

Chen et al., 2010 Chen T, Yu W-H, Izard J, Baranova OV, Lakshmanan A, Dewhirst FE. The Human Oral Microbiome Database: A web accessible resource for investigating oral microbe taxonomic and genomic information. *Database*. 2010;2010:baq013 **https://doi.org/10.1093/database/baq013**.

Cheng et al., 2022 Cheng L, Qi C, Yang H, et al. gutMGene: A comprehensive database for target genes of gut microbes and microbial metabolites. *Nucleic Acids Research*. 2022;50(D1):D795 **https://doi.org/10.1093/nar/gkab786**.

Cullen et al., 2020 Cullen CM, Aneja KK, Beyhan S, et al. Emerging priorities for microbiome research. *Frontiers in Microbiology*. 2020;11 **https://doi.org/10.3389/fmicb.2020.00136**.

Drula et al., 2022 Drula E, Garron M-L, Dogan S, Lombard V, Henrissat B, Terrapon N. The carbohydrate-active enzyme database: Functions and literature. *Nucleic Acids Research*. 2022;50(D1):D571 **https://doi.org/10.1093/nar/gkab1045**.

Everett et al., 2021 Everett C, Li C, Wilkinson JE, et al. Overview of the microbiome among nurses study (micro-N) as an example of prospective characterization of the microbiome within cohort studies. *Nature Protocols*. 2021;16(6):2724–2731 **https://doi.org/10.1038/s41596-021-00519-z**.

Field et al., 2008 Field D, Garrity G, Gray T, et al.... The minimum information about a genome sequence (MIGS) specification.

Nature Biotechnology. 2008;26(5):541–547 **https://doi.org/10.1038/nbt1360**.

Francavilla et al., 2023 Francavilla A, Ferrero G, Pardini B, et al. Gluten-free diet affects fecal small non-coding RNA profiles and microbiome composition in celiac disease supporting a host-gut microbiota crosstalk. *Gut Microbes*. 2023;15 **https://doi.org/10.1080/19490976.2023.2172955**.

Franzosa et al., 2014 Franzosa EA, Morgan XC, Segata N, et al. Relating the metatranscriptome and metagenome of the human gut. *Proceedings of the National Academy of Sciences of the United States of America*. 2014;111(22):E2329 **https://doi.org/10.1073/pnas.1319284111**.

Galperin et al., 2018 Galperin MY, Kristensen DM, Makarova KS, Wolf YI, Koonin EV. Microbial genome analysis: The COG approach. *Briefings in Bioinformatics*. 2018;20(4):1063–1070 **https://doi.org/10.1093/bib/bbx117**.

Gautam et al., 2023 Gautam A, Zeng W, Huson DH. MeganServer: Facilitating interactive access to metagenomic data on a server. *Bioinformatics (Oxford, England)*. 2023;39 **https://doi.org/10.1093/bioinformatics/btad105**.

Gerber et al., 2012 Gerber GK, Onderdonk AB, Bry L. Inferring. *PLoS Computational Biology*. 2012;8 **https://doi.org/10.1371/journal.pcbi.1002624**.

Ghazi et al., 2022 Ghazi AR, Münch PC, Chen D, Jensen J, Huttenhower C. *Strain identification and quantitative analysis in microbial communities*. *Journal of Molecular Biology*. 2022;434 **https://doi.org/10.1016/j.jmb.2022.167582** 167582.

Gloor et al., 2017 Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: And this is

not optional. *Frontiers in Microbiology*. 2017;8 **https://doi.org/10.3389/fmicb.2017.02224**.

Gonzalez et al., 2012 Gonzalez A, Stombaugh J, Lauber CL, Fierer N, Knight R. SitePainter: A tool for exploring biogeographical patterns. *Bioinformatics (Oxford, England)*. 2012;28(3):436–438 **https://doi.org/10.1093/bioinformatics/btr685**.

Grice et al., 2008 Grice EA, Kong HH, Renaud G, et al. A diversity profile of the human skin microbiota. *Genome Research*. 2008;18(7):1043–1050 **https://doi.org/10.1101/gr.075549.107**.

Griese et al., 2011 Griese M, Lange C, Soppa J. Ploidy in cyanobacteria. *FEMS Microbiology Letters*. 2011;323(2):124–131 **https://doi.org/10.1111/j.1574-6968.2011.02368.x**.

Guan et al., 2021 Guan H, Pu Y, Liu C, et al. Comparison of fecal collection methods on variation in gut metagenomics and untargeted metabolomics. *mSphere*. 2021;6 **https://doi.org/10.1128/msphere.00636-21**.

Guide for the care and use of laboratory animals, 2011 Guide for the care and use of laboratory animals. National Academies Press, National Academies Press, (2011).

Haas et al., 2012 Haas BJ, Chin M, Nusbaum C, Birren BW, Livny J. *How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes?*. *BMC Genomics*. 2012;13 **https://doi.org/10.1186/1471-2164-13-734** United States.

Haberer et al., 2005 Haberer G, Young S, Bharti AK, et al. Structure and architecture of the maize genome. *Plant Physiology*. 2005;139(4):1612–1624 **https://doi.org/10.1104/pp.105.068718**.

Handelsman et al., 1998 Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural

products. *Chemistry and Biology*. 1998;5(10):R245 **https://doi.org/10.1016/S1074-5521(98)90108-9**.

Heinken et al., 2023 Heinken A, Hertel J, Acharya G, et al. Genome-scale metabolic reconstruction of 7,302 human microorganisms for personalized medicine. *Nature Biotechnology*. 2023;41(9):1320–1331 **https://doi.org/10.1038/s41587-022-01628-0**.

Howe et al., 2021 Howe K, Dwinell M, Shimoyama M, et al.... The genome sequence of the Norway rat, Rattus norvegicus Berkenhout 1769. *Wellcome Open Research*. 2021;6:118 **https://doi.org/10.12688/wellcomeopenres.16854.1**.

Human Microbiome Project Consortium, 2012 Human Microbiome Project Consortium. A framework for human microbiome research. *Nature*. 2012;486(7402):215–221 **https://doi.org/10.1038/nature11209**.

Human Microbiome Project Consortium, 2012 Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*. 2012;486(7402):207–214 **https://doi.org/10.1038/nature11234**.

Hyde, 2023 Hyde K. Global consortium for the classification of fungi and fungus-like taxa. *Mycosphere*. 2023;14:1960–2012 **http://dx.doi.org/10.5943/mycosphere/14/1/23**.

Isokääntä et al., 2024 Isokääntä H, Pinto da Silva L, Karu N, et al. Comparative metabolomics and microbiome analysis of ethanol versus OMNImet/gene•GUT fecal stabilization. *Analytical Chemistry*. 2024;96(22):8893–8904 **https://doi.org/10.1021/acs.analchem.3c04436**.

Izard, 2015 Izard J. *Steps in Metagenomics: Let's Avoid Garbage in and Garbage Out*. *Metagenomics for Microbiology* United States:

Elsevier Inc.; 2015;1–23 **https://doi.org/10.1016/B978-0-12-410472-3.00001-4**.

Kanehisa et al., 2023 Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*. 2023;51(D1):D587 **https://doi.org/10.1093/nar/gkac963**.

Keefe et al., 2023 Keefe CR, Dillon MR, Gehret E, et al. Facilitating bioinformatics reproducibility with QIIME 2 Provenance Replay. *PLoS Computational Biology*. 2023;19(11):e1011676 **https://doi.org/10.1371/journal.pcbi.1011676**.

Kelly et al., 2015 Kelly BJ, Gross R, Bittinger K, et al. Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA. *Bioinformatics (Oxford, England)*. 2015;31(15):2461–2468 **https://doi.org/10.1093/bioinformatics/btv183**.

Kers and Saccenti, 2022 Kers JG, Saccenti E. The power of microbiome studies: Some considerations on which alpha and beta metrics to use and how to report results. *Frontiers in Microbiology*. 2022;12 **https://doi.org/10.3389/fmicb.2021.796025**.

Klappenbach et al., 2000 Klappenbach JA, Dunbar JM, Schmidt TM. rRNA operon copy number reflects ecological strategies of bacteria. *Applied and Environmental Microbiology*. 2000;66(4):1328–1333 **https://doi.org/10.1128/AEM.66.4.1328-1333.2000**.

Knight et al., 2018 Knight R, Vrbanac A, Taylor BC, et al.... Best practices for analysing microbiomes. *Nature Reviews Microbiology*. 2018;16(7):410–422 **https://doi.org/10.1038/s41579-018-0029-9**.

Komaki and Ishikawa, 2000 Komaki K, Ishikawa H. Genomic copy number of intracellular bacterial symbionts of aphids varies in

response to developmental stage and morph of their host. *Insect Biochemistry and Molecular Biology*. 2000;30(3):253–258 **https://doi.org/10.1016/s0965-1748(99)00125-3**.

Kumar et al., 2024 Kumar B, Lorusso E, Fosso B, Pesole G. A comprehensive overview of microbiome data in the light of machine learning applications: categorization, accessibility, and future directions. *Frontiers in Microbiology*. 2024;15 **https://doi.org/10.3389/fmicb.2024.1343572**.

Lange et al., 2022 Lange L, Berg G, Cernava T, et al. Microbiome ethics, guiding principles for microbiome research, use and knowledge management. *Environmental Microbiome*. 2022;17 **https://doi.org/10.1186/s40793-022-00444-y**.

Lloyd-Price et al., 2019 Lloyd-Price J, Arze C, Ananthakrishnan AN, et al.... Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*. 2019;569(7758):655–662 **https://doi.org/10.1038/s41586-019-1237-9**.

Martinez-Gutierrez and Aylward, 2022 Martinez-Gutierrez CA, Aylward FO. Genome size distributions in bacteria and archaea are strongly linked to evolutionary history at broad phylogenetic scales. *PLoS Genetics*. 2022;18 **https://doi.org/10.1371/journal.pgen.1010220**.

McDonald et al., 2024 McDonald D, Jiang Y, Balaban M, et al.... Greengenes2 unifies microbial data in a single reference tree. *Nature Biotechnology*. 2024;42(5):715–718 **https://doi.org/10.1038/s41587-023-01845-1**.

McFall-Ngai et al., 2013 McFall-Ngai M, Hadfield MG, Bosch TCG, et al.... Animals in a bacterial world, a new imperative for the life sciences. *Proceedings of the National Academy of Sciences*. 2013;110(9):3229–3236 **https://doi.org/10.1073/pnas.1218525110**.

Mei et al., 2024 Mei Z, Wang F, Bhosle A, et al.... Strain-specific gut microbial signatures in type 2 diabetes identified in a cross-cohort analysis of 8,117 metagenomes. *Nature Medicine*. 2024;30(8):2265–2276 **https://doi.org/10.1038/s41591-024-03067-7**.

Meier-Kolthoff et al., 2022 Meier-Kolthoff JP, Carbasse JS, Peinado-Olarte RL, Göker M. TYGS and LPSN: A database tandem for fast and reliable genome-based classification and nomenclature of prokaryotes. *Nucleic Acids Research*. 2022;50(1):D801 **https://doi.org/10.1093/nar/gkab902**.

Mirzayi et al., 2021 Mirzayi C, Renson A, Furlanello C, et al.... Reporting guidelines for human microbiome research: The STORMS checklist. *Nature Medicine*. 2021;27(11):1885–1892 **https://doi.org/10.1038/s41591-021-01552-x**.

Modlinska and Pisula, 2020 Modlinska K, Pisula W. The Norway rat, from an obnoxious pest to a laboratory pet. *eLife*. 2020;9 **https://doi.org/10.7554/elife.50651**.

Müller and Babel, 2003 Müller S, Babel W. Analysis of bacterial DNA patterns—An approach for controlling biotechnological processes. *Journal of Microbiological Methods*. 2003;55(3):851–858 **https://doi.org/10.1016/j.mimet.2003.08.003**.

Müller and Nebe-von-Caron, 2010 Müller S, Nebe-von-Caron G. Functional single-cell analyses: Flow cytometry and cell sorting of microbial populations and communities. *FEMS Microbiology Reviews*. 2010;34(4):554–587 **https://doi.org/10.1111/j.1574-6976.2010.00214.x**.

Collaboration, 2017 NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: A pooled analysis of 2416 population-based measurement studies in 128.9 million children,

adolescents, and adults. *Lancet*. 2017;390(10113):2627–2642 **https://doi.org/10.1016/S0140-6736(17)32129-3**.

Niederdorfer et al., 2017 Niederdorfer R, Besemer K, Battin TJ, Peter H. Ecological strategies and metabolic trade-offs of complex environmental biofilms. *npj Biofilms and Microbiomes*. 2017;3 **https://doi.org/10.1038/s41522-017-0029-y**.

Nishimura and Yoshizawa, 2022 Nishimura Y, Yoshizawa S. The OceanDNA MAG catalog contains over 50,000 prokaryotic genomes originated from various marine environments. *Scientific Data*. 2022;9 **https://doi.org/10.1038/s41597-022-01392-5**.

Odom et al., 2023 Odom AR, Faits T, Castro-Nallar E, Crandall KA, Johnson WE. Metagenomic profiling pipelines improve taxonomic classification for 16S amplicon sequencing data. *Scientific Reports*. 2023;13 **https://doi.org/10.1038/s41598-023-40799-x**.

Oren et al., 2023 Oren A, Arahal DR, Göker M, Moore ERB, Rossello-Mora R, Sutcliffe IC. International Code of Nomenclature of Prokaryotes Prokaryotic Code (2022 Revision). *International Journal of Systematic and Evolutionary Microbiology*. 2023;73 **https://doi.org/10.1099/ijsem.0.005585**.

Oren et al., 2021 Oren A, Arahal DR, Rosselló-Móra R, Sutcliffe IC, Moore ERB. Emendation of rules 5b, 8, 15 and 22 of the international code of nomenclature of prokaryotes to include the rank of phylum. *International Journal of Systematic and Evolutionary Microbiology*. 2021;71 **https://doi.org/10.1099/ijsem.0.004851**.

Oren and Garrity, 2021 Oren A, Garrity GM. Valid publication of the names of forty-two phyla of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology*. 2021;71 **https://doi.org/10.1099/ijsem.0.005056**.

Parte et al., 2020 Parte AC, Carbasse JS, Meier-Kolthoff JP, Reimer LC, Göker M. List of prokaryotic names with standing in nomenclature (LPSN) moves to the DSMZ. *International Journal of Systematic and Evolutionary Microbiology*. 2020;70(11):5607–5612 **https://doi.org/10.1099/ijsem.0.004332**.

Pecoraro et al., 2011 Pecoraro V, Zerulla K, Lange C, Soppa J, Aziz S. Quantification of ploidy in proteobacteria revealed the existence of monoploid, (mero-)oligoploid and polyploid species. *PLoS One*. 2011;6(1):e16392 **https://doi.org/10.1371/journal.pone.0016392**.

Rosa et al., 2012 Rosa PSL, Brooks JP, Deych E, et al. Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS One*. 2012;7(12):e52078 **https://doi.org/10.1371/journal.pone.0052078**.

Rosa et al., 2015 Rosa PSL, Zhou Y, Sodergren E, Weinstock G, Shannon WD. *Hypothesis Testing of Metagenomic Data. Metagenomics for Microbiology* United States: Elsevier Inc.; 2015;81–96 **https://doi.org/10.1016/B978-0-12-410472-3.00006-3**.

Russell and Burch, 1959 Russell WMS, Burch RL. *The principles of humane experimental technique. Animal Welfare* London: Methuen & Co. Limited; 1959;252.

Sayers et al., 2024 Sayers EW, Cavanaugh M, Clark K, et al. GenBank 2024 Update. *Nucleic Acids Research*. 2024;52(1):D134 **https://doi.org/10.1093/nar/gkad903**.

Schloss, 2020 Schloss PD. Reintroducing mothur: 10 years later. *Applied and Environmental Microbiology*. 2020;86 **https://doi.org/10.1128/AEM.02343-19**.

Schneider et al., 2017 Schneider VA, Graves-Lindsay T, Howe K, et al.... Evaluation of GRCh38 and de novo haploid genome

assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*. 2017;27(5):849–864 **https://doi.org/10.1101/gr.213611.116**.

Segata et al., 2012 Segata N, Haake SK, Mannon P, et al. *Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples*. *Genome Biology*. 2012;13 **https://doi.org/10.1186/gb-2012-13-6-r42** R42.

Segata et al., 2011 Segata N, Izard J, Waldron L, et al. Metagenomic biomarker discovery and explanation. *Genome Biology*. 2011;12 **https://doi.org/10.1186/gb-2011-12-6-r60**.

Stackebrandt, 2007 Stackebrandt E. Forces shaping bacterial systematics. *Microbe*. 2007;2(6):283–288 **https://doi.org/10.1128/microbe.2.283.1**.

Tannenbaum and Bennett, 2015 Tannenbaum J, Bennett BT. Russell and Burch's 3Rs then and now: The need for clarity in definition and purpose. *Journal of the American Association for Laboratory Animal Science*. 2015;54(2):120–132 **http://docserver.ingentaconnect.com/deliver/connect/aalas/15596109/v54n2/s2.pdf?expires=1428111103&id=81370426&titleid=72010024&accname=Elsevier+BV&checksum=233D2585271F9C34C5FF3A943724E0A7**.

Tarallo et al., 2022 Tarallo S, Ferrero G, De Filippis F, et al. Stool microRNA profiles reflect different dietary and gut microbiome patterns in healthy individuals. *Gut*. 2022;71(7):1302–1314 **https://doi.org/10.1136/gutjnl-2021-325168**.

Venter et al., 2001 Venter JC, Adams MD, Myers EW, et al.... The sequence of the human genome. *Science (New York, N.Y.)*. 2001;291(5507):1304–1351 **https://doi.org/10.1126/science.1058040**.

Verlinden et al., 2020 Verlinden H, Sterck L, Li J, et al. First draft genome assembly of the desert locust, Schistocerca gregaria. *F1000Research*. 2020;9:775 **https://doi.org/10.12688/f1000research.25148.2**.

Weber et al., 2018 Weber N, Liou D, Dommer J, et al. Nephele: A cloud platform for simplified, standardized and reproducible microbiome data analysis. *Bioinformatics (Oxford, England)*. 2018;34(8):1411–1413 **https://doi.org/10.1093/bioinformatics/btx617**.

World Health Organization, 2016 World Health Organization. *International Ethical Guidelines for Health-related Research Involving Humans* 4th ed. Switzerland: Council for International Organizations of Medical Sciences; 2016;.

Wetterstrand Wetterstrand, K.A. (2023). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). **https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data**.

Wu-Woods et al., 2023 Wu-Woods NJ, Barlow JT, Trigodet F, et al. Microbial-enrichment method enables high-throughput metagenomic characterization from host-rich samples. *Nature Methods*. 2023;20(11):1672–1682 **https://doi.org/10.1038/s41592-023-02025-4**.

Xiong et al., 2021 Xiong C, Singh BK, He J-Z, et al. Plant developmental stage drives the differentiation in ecological role of the maize microbiome. *Microbiome*. 2021;9 **https://doi.org/10.1186/s40168-021-01118-6**.

Yilmaz et al., 2011 Yilmaz P, Kottmann R, Field D, et al.... Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology*. 2011;29(5):415–420 **https://doi.org/10.1038/nbt.1823**.

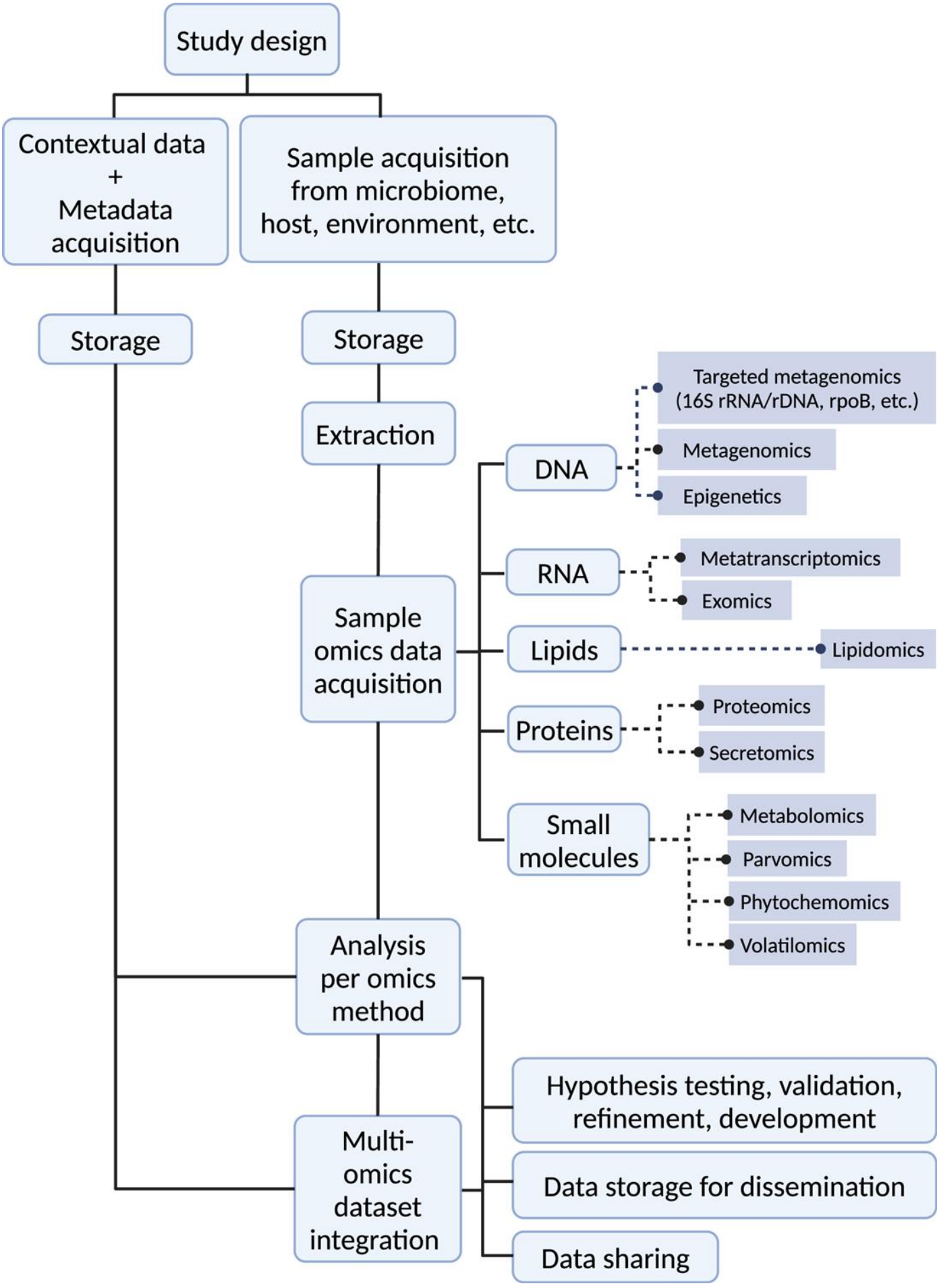
Zeeuwen et al., 2012 Zeeuwen PLJM, Boekhorst J, van den Bogaard EH, et al. Microbiome dynamics of human epidermis following skin barrier disruption. *Genome Biology*. 2012;13 **https://doi.org/10.1186/gb-2012-13-11-r101**.

Zhang et al., 2014 Zhang Z, Geng J, Tang X, et al. Spatial heterogeneity and co-occurrence patterns of human mucosal-associated intestinal microbiota. *The ISME Journal*. 2014;8(4):881–893 **https://doi.org/10.1038/ismej.2013.185**.

Zhou et al., 2019 Zhou W, Sailani MR, Contrepois K, et al.... Longitudinal multi-omics of host–microbe dynamics in prediabetes. *Nature*. 2019;569(7758):663–671 **https://doi.org/10.1038/s41586-019-1236-x**.

Zouiouich et al., 2023 Zouiouich S, Byrd DA, Hua X, et al. Stability of the fecal and oral microbiome over 2 years at −80°c for multiple collection methods. *Cancer Epidemiology, Biomarkers & Prevention*. 2023;32(3):444–451 **https://doi.org/10.1158/1055-9965.epi-22-0883**.

Description of Figure 1.1


Study design
Contextual data + Metadata acquisition
Sample acquisition from microbiome, host, environment, etc.
Storage
Storage
Extraction
Sample omics data acquisition
DNA
Targeted metagenomics (16S rRNA/rDNA, rpoB, etc.)
Metagenomics
Epigenetics
RNA
Metatranscriptomics
Exomics
Lipids
Lipidomics
Proteins
Proteomics
Secretomics
Small molecules
Metabolomics
Parvomics
Phytochemomics
Volatilomics
Analysis per omics method
Multi-omics dataset integration
Hypothesis testing, validation, refinement, development
Data storage for dissemination
Data sharing

The flowchart starts with 'Study design' which branches into 'Contextual data plus Metadata acquisition' and 'Sample acquisition from microbiome, host, environment, et cetera'. Both branches lead to 'Storage'. From 'Storage', the 'Sample acquisition' branch leads to 'Extraction', which then leads to 'Sample omics data acquisition'. This further branches into 'D N A', 'R N A', 'Lipids', 'Proteins' and 'Small molecules'. Each of these branches further subdivides: 'D N A' into 'Targeted metagenomics (16 S r R N A or r D N A, r p o B, et cetera)', 'Metagenomics' and 'Epigenetics'; 'R N A' into 'Metatranscriptomics' and 'Exomics'; 'Lipids' into 'Lipidomics'; 'Proteins' into 'Proteomics' and 'Secretomics'; 'Small molecules' into 'Metabolomics', 'Parvomics', 'Phytochemomics' and 'Volatilomics'. 'Sample omics data acquisition' and storage from contexual data plus metadata acquisition leads to 'Analysis per omics method', which then leads to 'Multi-omics dataset integration'. This further branches into 'Hypothesis testing, validation, refinement, development', 'Data storage for dissemination' and 'Data sharing'.

Navigate back to Figure 1.1

CHAPTER 2

The changing landscape of microbiome characterization methods: tools, trends, and trajectories

Naupaka Zimmerman[1], Emma Aronson[2] and Jacques Izard[3], [1]Department of Biology, University of San Francisco, San Francisco, CA, United States, [2]Department of Microbiology and Plant Pathology, University of California, Riverside, Riverside, CA, United States, [3]Department of Internal Medicine, University of Nebraska Medical Center, Omaha, NE, United States

Abstract

The techniques that microbial ecologists and environmental microbiologists use to study the complex microbial communities in environmental and host-associated contexts are ever changing. Here we provide context for the techniques that are currently in use and place them in a framework of the previous generation of analysis approaches. We include a state-of-the-art decision tree to help guide researchers through the considerations for choosing an approach when conducting microbiome characterization studies. Together we aim for these perspectives to give microbial ecologists and others studying host-associated and environmental microbiomes the context to enable them to choose the right methods to address their hypotheses.

Keywords

Microbiome; environmental metagenomics; historical comparison; host-associated microbiota; Illumina; Nanopore; PacBio; 454 pyrosequencing; decision tree; next-generation sequencing

Microbial metagenomics, metatranscriptomics, and associated analytic methodologies have seen rapid innovation and frequent changes. Over the past decade new and continually improving sequencing technologies and molecular approaches have made ever more complex analyses of microbial organisms possible, while also lowering costs for most projects. A challenge for a reader of the scientific literature is to understand some of the basic motivations for choosing one technology versus another, or why a particular technique leads to a certain type of result. It has been more than two decades since the seminal study by Venter et al. used shotgun sequencing of DNA from aquatic samples collected in the Sargasso Sea to characterize the community of microbes living there (Venter et al., 2004). About a decade into this new -omics era, we published a review with a decision tree to help unfamiliar researchers understand which of the extant sequence-based approaches to characterizing microbial communities could be used to best address their research questions (Zimmerman et al., 2014). At that time, RNA-focused analyses of microbial transcriptomes and other techniques that are now in common use were nascent. Here, a decade later, we provide an update on the available sequence-based analysis choices and explain the methodological transitions that have occurred and are currently in progress.

Taking a historical view of the technologies used in nucleic-acid-based "-omics" studies is crucial when engaging with the scientific literature. In addition to improved techniques for obtaining quality samples that accurately reflect the microbiome present at the time of Page 28 sampling, many recent advances have been driven by technological innovations in sequencing and bioinformatics tools. Historical perspective helps to:

- Contextualize past discoveries. None of the extant technologies can be used to answer all questions, but hypothesis generation and testing

over the last few decades have been led in part by the groundbreaking availability of high-dimensional datasets created via the use of complementary -omics studies (i.e., studies that combine comprehensive characterization of DNA, RNA, metabolites, and/or proteins from a given sample).

- Understand technological evolution over time and comprehend the technical terminologies associated with them, which can be somewhat confusing.
- Appreciate the limitations of older studies. Early metagenomics studies may have been limited by incomplete microbial diversity profiles due to lower sequencing depth or less advanced computational tools. Depending on the starting material, this concern may still be relevant to today's research.
- Evaluate methodologies to decide which to apply to a novel question. Knowing the technological backdrop helps when assessing whether study designs are comparable. Methods like amplicon sequencing (e.g., 16S rRNA gene sequencing) versus whole-genome shotgun sequencing have different strengths and limitations.
- Recognize improvements in data accuracy. Newer technologies provide higher-resolution data, thus leading to improvements in phylogenetic and functional assignments. That said, innovation continues apace and there is always room for improvement.
- Evaluate the reproducibility of findings. Changes in sample types, technology, depth of sequencing, and reference database(s) used, when comparing studies across different time periods, may lead to differences from previous studies or may complement previous work and thus contribute to a more comprehensive picture of a given biological phenomenon.

The vast literature surrounding microbial -omics analysis is rooted in the recognition that microbiomes are everywhere. They inhabit a wide range of

environments, from soil, oceans, and the atmosphere to the human body, animals, and plants, and even extreme environments like hydrothermal vents, deep subsurface brines, and polar ice caps. Beyond Page 29 Earth, our past space explorations may have left human and Earth microbiomes on the moon from equipment and other containers (NASA History Program Office, 2012), as well as circulating above us in the different space stations with long-term occupancy, including Mir (1986–2001) (Klintworth et al., 1999), the International Space Station (1998–present) (Simpson et al., 2023; Szydlowski et al., 2024), and the Tiangong space station (2021–present) (Zhang et al., 2024).

The recognition that microbiomes are omnipresent has had a profound impact on numerous fields of research. In human health, disruptions to these microbial communities have been linked to various diseases, including obesity, diabetes, and autoimmune disorders. This recognition has led to the popularization of probiotic supplements and the development of novel medications composed of bacterial communities (Benech et al., 2024; Sims et al., 2023). In the Salton Sea region of California, recent studies have shown that aerosols have a diverse microbial community and may be carrying bacterial lipo-polysaccharides that are harmful to respiratory health (Biddle et al., 2023; Freund et al., 2022). In agriculture, a diverse soil microbiome is increasingly recognized as a crucial component of soil health, where "healthy" soil communities can promote plant health and nutrient cycling and can even protect crops from disease (Dubey et al., 2019; Hartmann & Six, 2023). Some novel commercial seed products are sold with microbiome inocula added (Arnault et al., 2024). In the case of aquaculture, probiotics are sometimes included in commercial feed (Sumon et al., 2022). In biotechnology and the associated industry, microbial communities are increasingly being harnessed for applications in biofuel production, waste management, and bioremediation, where they help break down pollutants in the environment (Jurado et al., 2020; Ruan et al., 2024). In marine environments,

microbiomes influence global biogeochemical cycles, impacting climate and ecosystem health (King et al., 2023; Newton et al., 2011).

The way forward: traditional technologies scaled up and new -omics approaches

The transition from traditional microbiology to the integrated study of microbiomes as ecological communities represents a significant shift in how we study and understand microorganisms. The shift reflects a move from studying organisms in isolation to examining their behavior within complex communities in natural and host-associated Page 30 environments. This type of work previously existed (Kaplan et al., 2009; Muyzer, 1999), but most often the scientific hypotheses were focused on single organisms or simple communities with relatively low complexity and were limited both by cost and technical feasibility.

Two primary, cooccurring shifts were instrumental in the transition from traditional to -omics approaches for most microbial ecology research. First, there has generally been a decrease in the cost of molecular biology supplies and sequencing technologies. One product that is part of the breakthrough is the commercial availability of Hot Start Taq DNA polymerase for amplicon-based rRNA gene analysis of complex communities, decreasing the rate of chimera formation (Sharkey et al., 1994). Second, a greater acceptance that analyses of high-dimensional datasets is not only possible but is an essential approach for deciphering the biology of complex processes and phenomena. “High dimensional” in this case refers to the situation when the number of variables p is higher than the sample size n, that is, $p>n$, or any dataset which has thousands of features (variables), while having only dozens to hundreds of samples (instances). The variables used change across studies but could be taxa, genes, gene expression levels, proteins, metabolites, pathways, etc., while the instance is the number of samples collected for a particular project. This shift in analysis approach (from more traditional studies where $n>p$ to studies where $p>n$) necessitated the creation of new computational and

statistical tools for data processing and analysis (Abubucker et al., 2012; Beghini et al., 2021; Segata et al., 2011). For historical context, **Fig. 2.1** depicts a decision-making diagram from 2014 that encompasses the type of approaches that were then in use by researchers to manage the competing needs of a project, such as feasibility, time, outcomes, costs, interests, and scientific questions (Zimmerman et al., 2014).

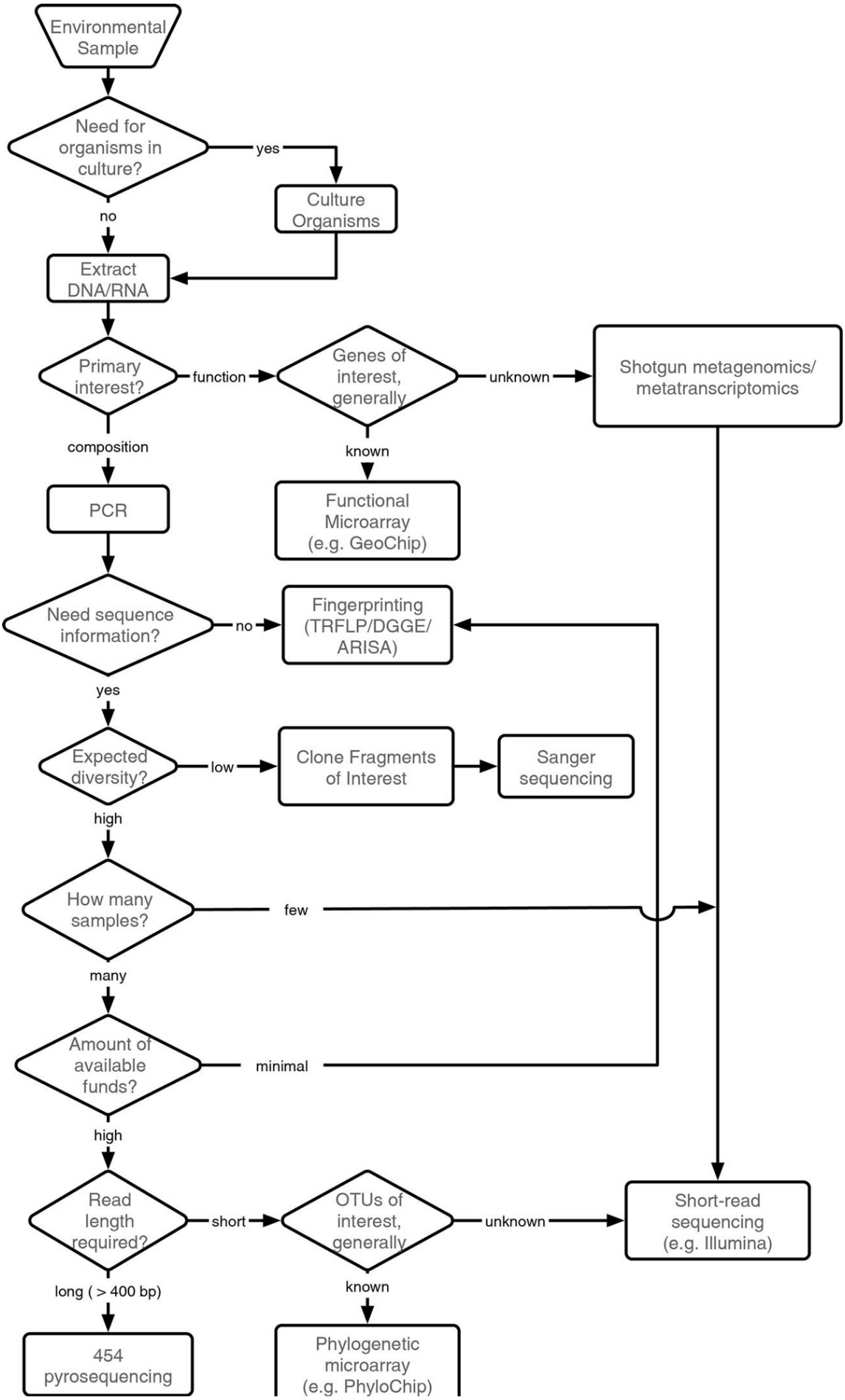

Environmental Sample
Need for organisms in culture?
yes
Culture Organisms
no
Extract DNA/RNA
Primary interest?
function
Genes of interest, generally
unknown
Shotgun metagenomics/ metatranscriptomics
composition
known
PCR
Functional Microarray (e.g. GeoChip)
Need sequence information?
no
Fingerprinting (TRFLP/DGGE/ ARISA)
yes
Expected diversity?
low
Clone Fragments of Interest
Sanger sequencing
high
How many samples?
few
many
Amount of available funds?
minimal
high
Read length required?
short
OTUs of interest, generally
unknown
Short-read sequencing (e.g. Illumina)
long (> 400 bp)
known
454 pyrosequencing
Phylogenetic microarray (e.g. PhyloChip)

FIGURE 2.1 2014 decision diagram for choosing a molecular approach for use in microbial ecology studies, originally published as Fig. 1 in Zimmerman et al. (2014) in *Frontiers in Ecology and the Environment*.

Follow for extended description

While a detailed discussion of the techniques used for sample collection and the methods to preserve samples prior to analysis are beyond the scope of this chapter, it is important to acknowledge that decisions in this area are driven by the anticipated use of the sample and associated data (Izard, 2015). One of the first decisions to make when planning sample collection is about whether the organism(s) will be cultivated. Single organism isolation techniques have been honed over the past century (Wolfgang et al., 2013), while isolation and cultivation of more complex consortia with, for example, microchambers designed via metagenome-based inference of physiological Page 31 Page 32 requirements is highly novel and have only been applied in roughly the last decade (Sizova et al., 2012). Once in culture, identification of single isolates or small consortia may rely on Sanger sequencing, most often of the 16S, ITS, or other so-called barcode loci (Das et al., 2014; Schoch et al., 2012).

However, for more comprehensive characterization, having the purified nucleic acids in hand is the start of all metagenomics and transcriptomics discovery processes. While not incompatible, the processes to obtain, from the same sample, analysis of composition (Who is there?), and of function (What can they potentially do?), may require different analytic and analysis strategies. As the field of microbial ecology has developed and expanded, more traditional approaches to the characterization of microbial communities (fragment length-based “fingerprinting” approaches) transitioned first to more sequencing-focused approaches (cloning and Sanger sequencing), and similarly early high-throughput approaches (microarrays) have given way to the more recent and the continued rise of ever-higher throughput next-

generation sequencing. This revolution began with sequencers from Solexa and Illumina, which offered higher throughput and accuracy but shorter read lengths (35–150 bp in either direction), and sequencers from Roche 454, which had comparatively lower throughput for the cost and had challenges with homopolymer regions but offered longer read lengths (400–1000 bp).

Early molecular approaches for characterizing microbial communities included techniques such as automated ribosomal intergenic spacer analysis, which characterizes microbial community diversity by using gel-based electrophoresis to assess length polymorphisms of the ribosomal RNA gene spacer region (Fisher & Triplett, 1999); denaturing gradient gel electrophoresis (DGGE), a technique that separates DNA fragments based on their melting behavior in a gel with a gradient of DNA denaturant (Muyzer et al., 1993); thermal gradient gel electrophoresis, which does the same with a thermal gradient (Benkato et al., 2017; Riesner et al., 1989); or terminal restriction fragment length polymorphism (TRFLP), a technique that involves PCR—amplifying a region of the rRNA gene, cutting the amplicons with restriction enzymes, and separating the resulting fragments by length to quantify microbial community compositional or functional diversity (Avaniss-Aghajani et al., 1994). None of these methods can identify unknown genetic variants at the nucleotide level within samples, but all can at least partially quantify the diversity of variants in a given sample.

Page 33 Molecular cloning and Sanger sequencing using universal primers for particular taxonomic groups (Pace et al., 1986; White et al., 1990; Woese, 1987) became quite popular in the 1990s and early 2000s and are occasionally still used for sequencing from mixed samples or communities. Sanger sequencing was often paired with fragment length and gradient gel methods to enable identification. None of these techniques represented major leaps forward in terms of sample throughput, but each allowed for a slightly different lens on the microbial community of interest. With all of these techniques, as more samples were added, more resources and time were required. Despite their limitations compared with modern high-throughput

sequencing methods, these technologies contributed to a dramatic expansion of our perception of microbial diversity.

The increase in rapid or high-throughput analysis of samples began with the broad availability of targeted microarrays in the 1990s. Microarray technology, wherein short probes of known sequences are fabricated onto a chip and DNA from a sample is allowed to hybridize with these probes, allows researchers to simultaneously examine thousands of microbial taxa (via barcode loci) or functional genes per sample. This capability made it possible to analyze complex microbial communities at a larger scale, identifying multiple species or gene functions without having to culture each organism in the laboratory. Microarray assays were relatively quick compared with traditional methods, and microarray-based methods were a more affordable option for the analysis of large datasets. Although they are less comprehensive than sequencing since the sequences they can identify are limited to those fabricated onto the chips at the time of their manufacture, microarrays provide a cost-effective solution for specific research questions in microbial ecology (Aronson et al., 2013). While this technique allowed for the high-throughput quantitation of known sequences in a sample in the absence of cultivation, the representation of novel genes or taxa was missing. Despite this limitation, microarrays enabled the study of the functional potential of complex microbial communities for the first time by targeting specific genes related to metabolic processes, resistance genes, or other functional markers. A range of gene chips was in use, including both phylogenetic—for example, PhyloChip, which targeted ribosomal gene sequences (DeSantis et al., 2007)—and functional microarrays—for example, GeoChip, which characterized genes involved in biogeochemical processes such as nutrient cycling (He et al., 2007). Page 34

The next big leap in throughput came with the advent of next-generation sequencing. While Sanger sequencing was generally limited to one gene per homogeneous sample at a time, next-generation sequencing provided the ablity to simultaneously sequence a single target gene from all organisms in a

community at once without cloning, or alternatively to carry out whole (meta)genome shotgun sequencing, with the option to multiplex multiple samples in a sequencing run via molecular barcoding before pooling. There were two commonly used high-throughput sequencing technologies in use when we published our earlier review: Roche's 454 pyrosequencing (Margulies et al., 2005) and Illumina's sequencing by synthesis (SBS) (Bentley et al., 2008). Roche's 454 technology was one of the early pioneers in next-generation sequencing. High cost and comparatively limited throughput opened the door to competitors like Illumina's SBS technology, which soon came to dominate the market due to its scalability, accuracy, and cost-effectiveness. Readers may have also encountered the names of other platforms such as SOLiD, Solexa, and Ion PGM from Life Technologies Corporation. Long-read sequencing, also called third-generation sequencing, started to appear on the market from companies like Oxford Nanopore with their ultraportable MinION sequencer (Branton et al., 2008) and Pacific Biosciences with their PacBio single-molecule real-time (SMRT) system (Izard, 2015).

The next decade sees shifting methodologies and an increased emphasis on large datasets

The change in prevalent technologies in scientific research is rarely abrupt. The analysis of data and the process of peer-review and publication take time, which blurs the transition to new technologies. Additionally, not all legacy technologies are inherently limited, nor are they necessarily static. Existing techniques, when modified with new reagents, might continue to provide results of increasing value. Sanger sequencing is by no means obsolete; it is just currently reserved for niche applications, such as assessing single PCR-amplified loci in isolated organisms. Fingerprinting has also become a niche application. While lacking phylogenetic and functional attribution, fingerprinting can still be practical for the selection of samples in longitudinal data when a shift is expected but the timing is unknown. That said, the

fragment-length, cloning-based, and gene chip techniques described above have been almost entirely supplanted with high-throughput sequencing approaches. In **Fig. 2.2** we focus on presently dominant
Page 35 Page 36
technologies and provide a decision tree to help guide researchers aiming to select an appropriate technique.

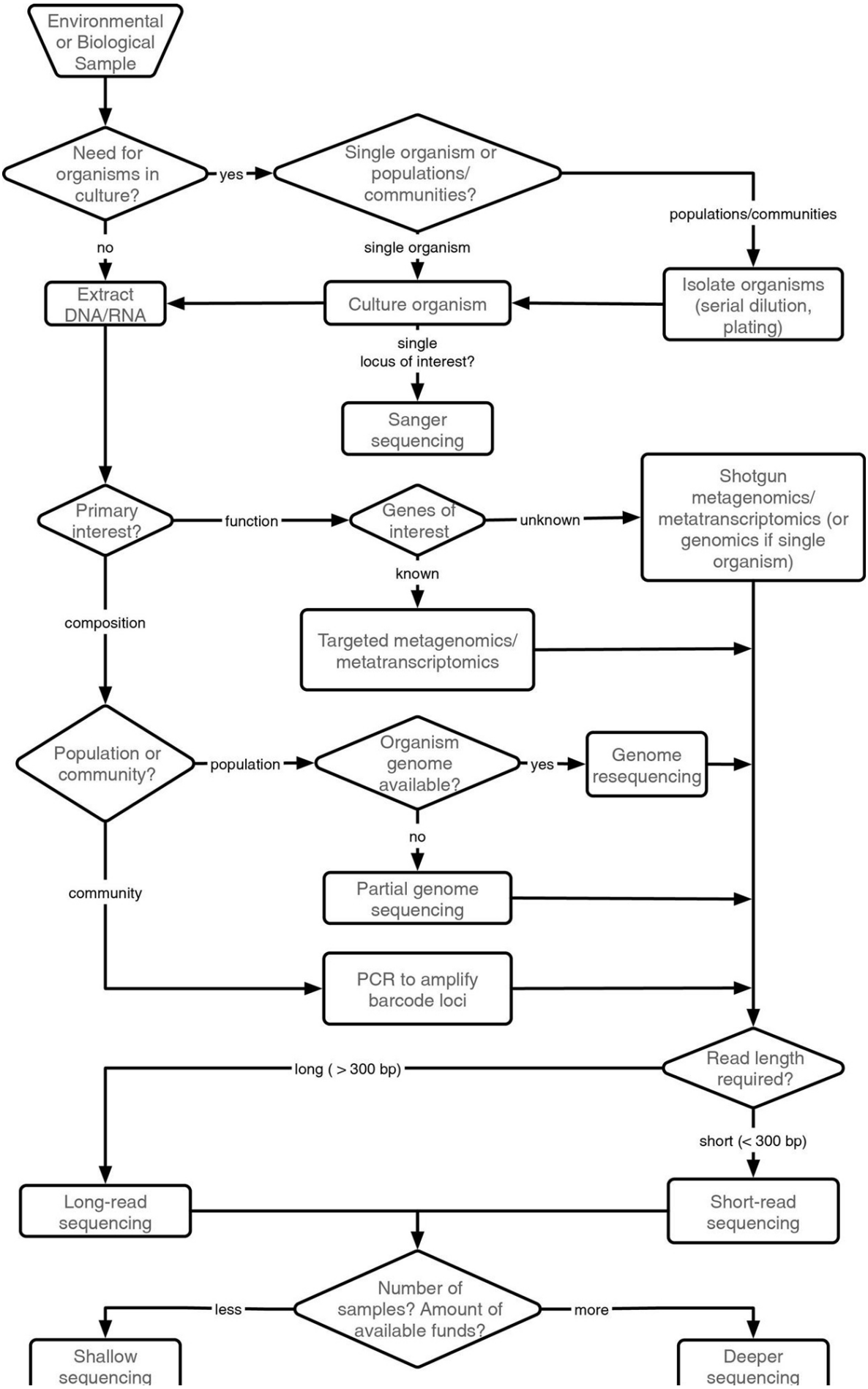

Environmental or Biological Sample
Need for organisms in culture?
yes
Single organism or populations/ communities?
populations/communities
no
single organism
Extract DNA/RNA
Culture organism
Isolate organisms (serial dilution, plating)
single locus of interest?
Sanger sequencing
Primary interest?
function
Genes of interest
unknown
Shotgun metagenomics/ metatranscriptomics (or genomics if single organism)
known
composition
Targeted metagenomics/ metatranscriptomics
Population or community?
population
Organism genome available?
yes
Genome resequencing
no
community
Partial genome sequencing
PCR to amplify barcode loci
Read length required?
long (> 300 bp)
short (< 300 bp)
Long-read sequencing
Short-read sequencing
Number of samples? Amount of available funds?
less
more
Shallow sequencing
Deeper sequencing

FIGURE 2.2 Updated decision diagram to reflect changes in techniques and approaches since 2014.

Follow for extended description

In terms of changes that have occurred within high-throughput sequencing technologies and approaches, Roche's 454 pyrosequencing is no longer in use, nor is it commercially available. Illumina still uses sequencing-by-synthesis and read lengths are roughly similar (<300 bp when combining forward and reverse reads) to ones developed 10 years ago, but over the course of the past 10 years, the cost per base has continued to drop and the amount of sequence data generated per run has continued to increase, which has enabled much deeper levels of sequencing per sample. Another trend has been new instruments and reagent kits that offer a reduction in sequencing time for a given amount of data produced. Higher sequencing costs a decade ago often led studies to use more targeted metagenomics or metatranscriptomic approaches, for example by using capture probes or specific primers (Suenaga, 2012). Falling sequencing costs, particularly for the highest-output sequencers such as Illumina's NovaSeq, now make it increasingly tractable for most laboratories to conduct deep shotgun sequencing of even complex environmental microbial communities in the style of seminal studies like Venter et al. (2004). Over the last decade there has been a proliferation of different sequencing instruments and library preparation kits that allow for a broad range of sequencing depths and applications. On the smaller end in terms of sequencing depth, in the last few years Illumina has introduced a large number of new sequencing instruments and library preparation and sequencing reagent kits. This has led some per-run sequencing costs to start to rise again, although per-base costs in the highest throughput options either (depending on the reagent and instrument) continue to fall or else to remain relatively stagnant. Newer single-molecule long-read sequencing approaches like nanopore sequencing from Oxford Nanopore (Branton et al., 2008) and SMRT sequencing from Pacific Biosciences (Levene et al., 2003) are now

widely available and can produce reads from tens of kilobases up to reads in excess of a megabase. These also allow for direct sequencing of some epigenetic markers like methylated bases (Flusberg et al., 2010; Rand et al., 2017; Simpson et al., 2017). The other trend of note, as alluded to above in the case of Illumina, is that sequencing companies have substantially increased the number of different sequencing hardware models, segmenting the market for new sequencers by producing instruments that span a wide range of price/output levels, from low capital Page 37 investment and relatively low sequence output capacity to high capital investment and relatively high sequence output capacity.

Many recent (meta)genome studies integrate across these different long-read and short-read techniques to enhance assembly, particularly those focused on de novo eukaryotic genome sequencing or complex eukaryotic metagenomes or metatranscriptomes (Frank et al., 2016; Koren et al., 2012; Pendleton et al., 2015). Similar approaches are used to finish draft prokaryotic genomes (Luan et al., 2024; Wick et al., 2022). In contrast, studies of organisms with well-established reference genomes where de novo assembly is not required typically rely solely on short-read approaches for re-sequencing.

Another large change in microbiome studies is the focal shift from characterizing the diversity and makeup of microbial communities to studies that work to infer function based on the sequencing of functional genes through metagenomic approaches. However, for characterizing host-associated or environmental microbial communities, amplicon sequencing of target barcode loci (e.g., 16S rRNA gene for bacteria or ITS for fungi) is still in wide use. As sequencing costs continue to fall some studies are now using shotgun sequencing and new informatics approaches [e.g., MetaPhlan (Blanco-Míguez et al., 2023), one of the existing computing pipelines] to enable the quantification of community composition and function, as these types of approaches can be less biased than more targeted amplicon methods. Shotgun sequencing for the purpose of conducting metagenomic or

metatranscriptomic studies has become increasingly tractable, not only because of falling sequencing costs and increased sequencing depth but also because computational capacity has continued to increase, which newly enables de novo assembly of increasingly complex samples.

For PCR and amplicon sequencing-based approaches in host-associated microbiomes, there have been both molecular advances and computational advances. Peptide nucleic acids are now routinely used to block the amplification of host DNA to preferentially amplify microbial templates (Ørum et al., 1993; von Wintzingerode et al., 2000). This is particularly useful for blocking the amplification of mitochondrial and chloroplast genomic DNA when studying eukaryotic host-associated bacteria. This approach was less commonly used a decade ago. As sequence read quality of high-throughput approaches has continued to increase, the informatics used to analyze microbial communities has also changed. While in 2014 clustering of amplicon reads into operational taxonomic units (OTUs, similar sequences Page 38 grouped into clusters approximately representing a taxonomic unit like a species or genus) was the standard approach, currently many community diversity studies focus on the analysis of amplicon sequence variants, which are roughly the equivalent of OTUs clustered at 100% similarity. This has been driven by improvements in denoising algorithms that can statistically account for the errors introduced by PCR and sequencing (Callahan et al., 2016; Edgar & Flyvbjerg, 2015; Edgar, 2016; Nearing et al., 2018; Nilsen et al., 2024). This shift in approach allows for analyses focused on within-species diversity rather than only looking at groupings that approximate higher taxonomic levels.

Looking forward: expectations and predictions

The concept of the microbiome is now well established both in the scientific community and as a common term in newspapers and magazines when describing environmental and public health-related topics. Looking forward,

and based on the advances and direction of research aims described above, we predict a set of research developments in the coming years.

Longitudinal analysis will become more common. Now that sequencing technologies are well defined, and throughput is high for some providers, the sequencing step is no longer the limiting factor. Yet, collection and analysis, the steps before and after sequencing, are still challenging. We still have a limited number of studies taking place over long time periods. As sample preparation and preservation techniques continue to evolve, we should look forward to greater data availability from studies characterizing microbiome biology along a timeline, an intervention, or a natural cycle.

Multisite collection (i.e., repeated sampling across different sites on a host) will become easier to interpret as new analysis tools beyond just visualization will start to appear. While the Human Microbiome Project pioneered multisite analysis in human populations (Segata et al., 2012), many studies have sampled the microbiome at multiple locations in/on plants and animals. The bottlenecks are still the integration of function and location (Amend et al., 2022; Li et al., 2023). This task gets even more complicated when the collection reaches high sample numbers, such as tens of thousands of samples (Everett et al., 2021).

Better management of "huge" raw datasets to reach tables can be included in statistical analysis. While sequencing centers can rapidly provide researchers with terabytes of data, data management is still a challenging endeavor for most scientists. We look forward to new tools focusing on easing the process while assuring accuracy and capturing data provenance. Relatedly, modeling microbiome biology over time is still in its infancy. Computing advances and the availability of numerous deep metagenomic datasets will usher in an expansion in mathematical modeling and computational tools being applied to environmental and host-associated microbiomes (Maringanti et al., 2022).

Standardization of sampling and analysis pipelines to enable comparison may become more widespread. An example for human health is the need to

standardize stool microbiome analysis provided by direct-to-consumer microbiome testing services. If the goal of these providers is to provide decision-making data for guiding symptom treatment or nutrition, the participant should have the same results wherever they go (Hoffmann et al., 2024). The same is true for environmental samples used to compare microbiomes from diverse ecosystems. There has been some community momentum behind protocols such as those developed by the Earth Microbiome Project (Shaffer et al., 2022); however the broad applicability of such short amplicon sequencing may be limited.

Integrated multiomics approaches will be more common. Functional genomics, proteomics, and metabolomics need to be better integrated into microbiome studies to provide a comprehensive view of microbial activity and its impacts (Chetty & Blekhman, 2024). The range of technological and analytical advances here described all contribute to decreasing cost and increasing ease of integration across these datasets. Furthermore, *variation in biological systems* caused by cyclical factors requires integration of understanding across multiple scales of a biological system. From the change in the population of a microbiome at a location, the biogeography and associated phylogenetic construction of the community, the functional changes due to gene expression, the fluctuations in the transcriptome and proteome, the small and large molecules produced in response to environmental and nutritional cues, and the changes in metabolomes and lipidomes, all of these components change in response to diurnal and seasonal factors. While characterization of these changes is now possible, our understanding of their implications and mechanisms is still nascent.

Along those lines, we should go back to the basics to directly test the function of genes beyond deduction by similarities. Computational efforts can only go so far; robust knowledge of gene function requires a return to the “wet lab” (involving directly the biological agent of interest in laboratory setting) to better understand gene/function relationships. Scaling up gene/function identification and validation is far from a simple task and

will require substantial effort and funding to ensure progress. Recent advances may allow for the culture in consortia of a wider range of microorganisms that resist pure culture, expanding our ability to generate functional profiles.

The increased use of artificial intelligence (AI) will have a large impact on our ability to analyze the overwhelming amount of microbiome data available. While some 'AI' approaches have been used since the inception of metagenomics, including, for example, machine learning methods like ordination, the recent explosion of work on deep neural networks will certainly increasingly begin to intersect with microbial ecology research. However, to provide mechanistic scientific insights, new generations of AI models will need to be interpretable and move away from proprietary “black box” systems. The health of humans, animals, crops, and other organisms will depend on how the data are analyzed.

Finally, we expect an increase in the number of commercialized products linked to microbiome studies and discoveries. Examples of success are the first FDA-approved human microbiome-based treatments (Benech et al., 2024). Initially we anticipate a wide array of microbiome products targeted at humans across the realms of nonclinical probiotics and prebiotic supplements, eventually transitioning to approved clinical treatments based on microbiome products. As is apparent in these nascent human-focused industries, we expect similar success in the coming years in aquaculture and agriculture (Brugman et al., 2018).

Conclusions

Although it is challenging to predict the next decade’s advancements in sequencing and analysis techniques, it is likely that costs will continue to decrease while quality, quantity, and read lengths will continue to improve. These advancements will likely drive a shift toward studies that explore and test hypotheses using data from extensive de novo metagenomes and metatranscriptomes, rather than relying on more limited approaches like

amplicon sequencing. As a result, there will be a growing demand for more advanced computing resources and bioinformatics tools. However, the real challenge may lie in the Page 41 annotation and interpretation of the increasingly rich datasets we are generating. As we look forward to embracing the potential of these powerful technologies, it is equally important to look back and remember the value of traditional natural history and culture techniques, which, when combined with new approaches, will help us continue to deepen our understanding of the myriad microbial communities in and around us.

Acknowledgments

We are grateful to Christian G. Klatt and Jizhong Zhou who coauthored the 2014 review published in *Frontiers in Ecology and the Environment*.

References

Abubucker et al., 2012 Abubucker S, Segata N, Goll J, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Computational Biology*. 2012;8(6):e1002358 **https://doi.org/10.1371/journal.pcbi.1002358**.

Amend et al., 2022 Amend AS, Swift SOI, Darcy JL, et al.... A ridge-to-reef ecosystem microbial census reveals environmental reservoirs for animal and plant microbiomes. *Proceedings of the National Academy of Sciences of the United States of America*. 2022;119 **https://doi.org/10.1073/pnas.2204146119**.

Arnault et al., 2024 Arnault G, Marais C, Préveaux A, et al. Seedling microbiota engineering using bacterial synthetic community inoculation on seeds. *FEMS Microbiology Ecology*. 2024;100 **https://doi.org/10.1093/femsec/fiae027**.

Aronson et al., 2013 Aronson EL, Dubinsky EA, Helliker BR. Effects of nitrogen addition on soil microbial diversity and methane cycling capacity depend on drainage conditions in a pine forest soil. *Soil Biology and Biochemistry*. 2013;62:119–128 **https://doi.org/10.1016/j.soilbio.2013.03.005**.

Avaniss-Aghajani et al., 1994 Avaniss-Aghajani E, Jones K, Chapman D, Brunk C. A molecular technique for identification of bacteria using small subunit ribosomal RNA sequences. *Biotechniques*. 1994;17(1):144–149.

Beghini et al., 2021 Beghini F, McIver LJ, Blanco-Míguez A, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife*. 2021;10 **https://doi.org/10.7554/elife.65088**.

Benech et al., 2024 Benech N, Barbut F, Fitzpatrick F, et al. Update on microbiota-derived therapies for recurrent Clostridioides difficile infections. *Clinical Microbiology and Infection*. 2024;30(4):462–468 **https://doi.org/10.1016/j.cmi.2023.12.007**.

Benkato et al., 2017 Benkato K, O'Brien B, Bui MN, Jasinski DL, Guo P, Khisamutdinov EF. Evaluation of thermal stability of RNA nanoparticles by temperature gradient gel electrophoresis (TGGE) in native condition. *Methods in Molecular Biology*. 2017;1632:123–133 **https://doi.org/10.1007/978-1-4939-7138-1_8**.

Bentley et al., 2008 Bentley DR, Balasubramanian S, Swerdlow HP, et al.... Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008;456(7218):53–59 **https://doi.org/10.1038/nature07517**.

Biddle et al., 2023 Biddle TA, Yisrael K, Drover R, et al. Aerosolized aqueous dust extracts collected near a drying lake trigger acute neutrophilic pulmonary inflammation reminiscent of microbial innate immune ligands. *Science of The Total Environment*.

2023;858(Pt 3):159882 **https://doi.org/10.1016/j.scitotenv.2022.159882**.

Blanco-Míguez et al., 2023 Blanco-Míguez A, Beghini F, Cumbo F, et al.... Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nature Biotechnology*. 2023;41(11):1633–1644 **https://doi.org/10.1038/s41587-023-01688-w**.

Branton et al., 2008 Branton D, Deamer DW, Marziali A, et al.... The potential and challenges of nanopore sequencing. *Nature Biotechnology*. 2008;26(10):1146–1153 **https://doi.org/10.1038/nbt.1495**.

Brugman et al., 2018 Brugman S, Ikeda-Ohtsubo W, Braber S, Folkerts G, Pieterse CMJ, Bakker PAHM. A comparative review on microbiota manipulation: Lessons from fish, plants, livestock, and human research. *Frontiers in Nutrition*. 2018;5 **https://doi.org/10.3389/fnut.2018.00080**.

Callahan et al., 2016 Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*. 2016;13(7):581–583 **https://doi.org/10.1038/nmeth.3869**.

Chetty and Blekhman, 2024 Chetty A, Blekhman R. Multi-omic approaches for host-microbiome data integration. *Gut Microbes*. 2024;16 **https://doi.org/10.1080/19490976.2023.2297860**.

Das et al., 2014 Das S, Dash HR, Mangwani N, Chakraborty J, Kumari S. Understanding molecular identification and polyphasic taxonomic approaches for genetic relatedness and phylogenetic relationships of microorganisms. *Journal of Microbiological Methods*. 2014;103:80–100 **https://doi.org/10.1016/j.mimet.2014.05.013**.

DeSantis et al., 2007 DeSantis TZ, Brodie EL, Moberg JP, Zubieta IX, Piceno YM, Andersen GL. High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. *Microbial Ecology*. 2007;53(3):371–383 **https://doi.org/10.1007/s00248-006-9134-9**.

Dubey et al., 2019 Dubey A, Malla MA, Khan F, et al. Soil microbiome: A key player for conservation of soil health under changing climate. *Biodiversity and Conservation*. 2019;28(8-9):2405–2429 **https://doi.org/10.1007/s10531-019-01760-5**.

Edgar, 2016 Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv* 2016; **https://doi.org/10.1101/081257**.

Edgar and Flyvbjerg, 2015 Edgar RC, Flyvbjerg H. Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics (Oxford, England)*. 2015;31(21):3476–3482 **https://doi.org/10.1093/bioinformatics/btv401**.

Everett et al., 2021 Everett C, Li C, Wilkinson JE, et al. Overview of the microbiome among nurses study (Micro-N) as an example of prospective characterization of the microbiome within cohort studies. *Nature Protocols*. 2021;16(6):2724–2731 **https://doi.org/10.1038/s41596-021-00519-z**.

Fisher and Triplett, 1999 Fisher MM, Triplett EW. Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. *Applied and Environmental Microbiology*. 1999;65(10):4630–4636 **https://doi.org/10.1128/aem.65.10.4630-4636.1999**.

Flusberg et al., 2010 Flusberg BA, Webster DR, Lee JH, et al. Direct detection of DNA methylation during single-molecule, real-time

sequencing. *Nature Methods*. 2010;7(6):461–465 **https://doi.org/10.1038/nmeth.1459**.

Frank et al., 2016 Frank JA, Pan Y, Tooming-Klunderud A, et al. Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data.. *Scientific Reports*. 2016;6:25373 **https://doi.org/10.1038/srep25373**.

Freund et al., 2022 Freund H, Maltz M, Swenson M, et al. Microbiome interactions and their ecological implications at the Salton Sea. *California Agriculture*. 2022;76(1):16–26 **https://doi.org/10.3733/ca.2022a0002**.

Hartmann and Six, 2023 Hartmann M, Six J. Soil structure and microbiome functions in agroecosystems. *Nature Reviews Earth and Environment*. 2023;4(1):4–18 **https://doi.org/10.1038/s43017-022-00366-w**.

He et al., 2007 He Z, Gentry TJ, Schadt CW, et al. GeoChip: A comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *The ISME Journal*. 2007;1(1):67–77 **https://doi.org/10.1038/ismej.2007.2**.

Hoffmann et al., 2024 Hoffmann DE, von Rosenvinge EC, Roghmann MC, Palumbo FB, McDonald D, Ravel J. The DTC microbiome testing industry needs more regulation. *Science (New York, N.Y.)*. 2024;383(6688):1176–1179 **https://doi.org/10.1126/science.adk4271**.

Izard, 2015 Izard J. *Steps in Metagenomics: Let's Avoid Garbage in and Garbage Out*. *Metagenomics for Microbiology* United States: Elsevier Inc.; 2015;1–23 **https://doi.org/10.1016/B978-0-12-410472-3.00001-4**.

Jurado et al., 2020 Jurado MM, Camelo-Castillo AJ, Suárez-Estrella F, et al. Integral approach using bacterial microbiome to stabilize municipal solid waste. *Journal of Environmental Management*.

2020;265 **https://doi.org/10.1016/j.jenvman.2020.110528**.

Kaplan et al., 2009 Kaplan CW, Lux R, Haake SK, Shi W. The Fusobacterium nucleatum outer membrane protein RadD is an arginine-inhibitable adhesin required for inter-species adherence and the structured architecture of multispecies biofilm. *Molecular Microbiology*. 2009;71(1):35–47 **https://doi.org/10.1111/j.1365-2958.2008.06503.x**.

King et al., 2023 King NG, Wilmes SB, Browett SS, et al. Seasonal development of a tidal mixing front drives shifts in community structure and diversity of bacterioplankton. *Molecular Ecology*. 2023;32(18):5201–5210 **https://doi.org/10.1111/mec.17097**.

Klintworth et al., 1999 Klintworth R, Reher HJ, Viktorov AN, Bohle D. Biological induced corrosion of materials II: New test methods and experiences from mir station. *Acta Astronautica*. 1999;44(7-12):569–578 **https://doi.org/10.1016/s0094-5765(99)00069-7**.

Koren et al., 2012 Koren S, Schatz MC, Walenz BP, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*. 2012;30(7):693–700 **https://doi.org/10.1038/nbt.2280**.

Levene et al., 2003 Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science (New York, N.Y.)*. 2003;299(5607):682–686 **https://doi.org/10.1126/science.1079700**.

Li et al., 2023 Li J, Bates KA, Hoang KL, Hector TE, Knowles SCL, King KC. Experimental temperatures shape host microbiome diversity and composition. *Global Change Biology*. 2023;29(1):41–56 **https://doi.org/10.1111/gcb.16429**.

Luan et al., 2024 Luan T, Commichaux S, Hoffmann M, et al. Benchmarking short and long read polishing tools for nanopore assemblies: Achieving near-perfect genomes for outbreak isolates. *BMC Genomics*. 2024;25 **https://doi.org/10.1186/s12864-024-10582-x**.

Margulies et al., 2005 Margulies M, Egholm M, Altman WE, et al.... Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 2005;437(7057):376–380 **https://doi.org/10.1038/nature03959**.

Maringanti et al., 2022 Maringanti VS, Bucci V, Gerber GK. MDITRE: Scalable and interpretable machine learning for predicting host status from temporal microbiome dynamics. *mSystems*. 2022;7 **https://doi.org/10.1128/msystems.00132-22**.

Muyzer, 1999 Muyzer G. DGGE/TGGE a method for identifying genes from natural ecosystems. *Current Opinion in Microbiology*. 1999;2(3):317–322 **https://doi.org/10.1016/s1369-5274(99)80055-1**.

Muyzer et al., 1993 Muyzer G, De Waal EC, Uitterlinden AG. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Applied and Environmental Microbiology*. 1993;59(3):695–700 **https://doi.org/10.1128/aem.59.3.695-700.1993**.

NASA History Program Office, 2012 NASA History Program Office, Catalogue of Manmade Material on the Moon. (2012).

Nearing et al., 2018 Nearing JT, Douglas GM, Comeau AM, Langille MGI. Denoising the denoisers: An independent evaluation of microbiome sequence error-correction approaches. *PeerJ*. 2018;6 **https://doi.org/10.7717/peerj.5364**.

Newton et al., 2011 Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S. A guide to the natural history of freshwater lake bacteria. *Microbiology and Molecular Biology Reviews*. 2011;75(1):14–49 **https://doi.org/10.1128/MMBR.00028-10**.

Nilsen et al., 2024 Nilsen T, Snipen L-G, Angell IL, et al. Swarm and UNOISE outperform DADA2 and Deblur for denoising high-diversity marine seafloor samples. *ISME Communications*. 2024;4 **https://doi.org/10.1093/ismeco/ycae071**.

Ørum et al., 1993 Ørum H, Nielsen PE, Egholm M, Berg RH, Buchardt O, Stanley C. Single base pair mutation analysis by PNA directed PCR clamping. *Nucleic Acids Research*. 1993;21(23):5332–5336 **https://doi.org/10.1093/nar/21.23.5332**.

Pace et al., 1986 Pace NR, Stahl DA, Lane DJ, Olsen GJ. *The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences* Springer Science and Business Media LLC 1986;1–55 **https://doi.org/10.1007/978-1-4757-0611-6_1**.

Pendleton et al., 2015 Pendleton M, Sebra R, Pang AWC, et al.... Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods*. 2015;12(8):780–786 **https://doi.org/10.1038/nmeth.3454**.

Rand et al., 2017 Rand AC, Jain M, Eizenga JM, et al. Mapping DNA methylation with high-throughput nanopore sequencing. *Nature Methods*. 2017;14(4):411–413 **https://doi.org/10.1038/nmeth.4189**.

Riesner et al., 1989 Riesner D, Steger G, Zimmat R, et al. Temperature-gradient gel electrophoresis of nucleic acids: Analysis of conformational transitions, sequence variations, and protein-nucleic acid interactions. *Electrophoresis*. 1989;10(5-6):377–389 **https://doi.org/10.1002/elps.1150100516**.

Ruan et al., 2024 Ruan Z, Chen K, Cao W, et al. Engineering natural microbiomes toward enhanced bioremediation by microbiome modeling. *Nature Communications*. 2024;15 **https://doi.org/10.1038/s41467-024-49098-z**.

Schoch et al., 2012 Schoch CL, Seifert KA, Huhndorf S, et al.... Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America*. 2012;109(16):6241–6246 **https://doi.org/10.1073/pnas.1117018109**.

Segata et al., 2012 Segata N, Haake SK, Mannon P, et al. *Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples*. *Genome Biology*. 2012;13 **https://doi.org/10.1186/gb-2012-13-6-r42** R42.

Segata et al., 2011 Segata N, Izard J, Waldron L, et al. Metagenomic biomarker discovery and explanation. *Genome Biology*. 2011;12 **https://doi.org/10.1186/gb-2011-12-6-r60**.

Shaffer et al., 2022 Shaffer JP, Nothias LF, Thompson LR, et al.... Standardized multi-omics of Earth's microbiomes reveals microbial and metabolite diversity. *Nature Microbiology*. 2022;7(12):2128–2150 **https://doi.org/10.1038/s41564-022-01266-x**.

Sharkey et al., 1994 Sharkey DJ, Scalice ER, Christy KG, Atwood SM, Daiss JL. Antibodies as thermolabile switches: High temperature triggering for the polymerase chain reaction. *Bio/Technology*. 1994;12(5):506–509 **https://doi.org/10.1038/nbt0594-506**.

Simpson et al., 2023 Simpson AC, Sengupta P, Zhang F, et al. Phylogenomics, phenotypic, and functional traits of five novel (Earth-derived) bacterial species isolated from the International Space Station and their prevalence in metagenomes. *Scientific*

Reports. 2023;13 **https://doi.org/10.1038/s41598-023-44172-w**.

Simpson et al., 2017 Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods*. 2017;14(4):407–410 **https://doi.org/10.1038/nmeth.4184**.

Sims et al., 2023 Sims MD, Khanna S, Feuerstadt P, et al.... Safety and tolerability of SER-109 as an investigational microbiome therapeutic in adults with recurrent clostridioides difficile infection a phase 3, open-label, single-arm trial. *JAMA Network Open*. 2023;6 **https://doi.org/10.1001/jamanetworkopen.2022.55758**.

Sizova et al., 2012 Sizova MV, Hohmann T, Hazen A, et al. New approaches for isolation of previously uncultivated oral bacteria. *Applied and Environmental Microbiology*. 2012;78(1):194–203 **https://doi.org/10.1128/aem.06813-11**.

Suenaga, 2012 Suenaga H. Targeted metagenomics: A high-resolution metagenomics approach for specific gene clusters in complex microbial communities. *Environmental Microbiology*. 2012;14(1):13–22 **https://doi.org/10.1111/j.1462-2920.2011.02438.x**.

Sumon et al., 2022 Sumon MAA, Sumon TA, Hussain MA, et al. Single and multi-strain probiotics supplementation in commercially prominent finfish aquaculture: Review of the current knowledge. *Journal of Microbiology and Biotechnology*. 2022;32(6):681–698 **https://doi.org/10.4014/jmb.2202.02032**.

Szydlowski et al., 2024 Szydlowski LM, Bulbul AA, Simpson AC, et al. Adaptation to space conditions of novel bacterial species isolated from the International Space Station revealed by functional gene annotations and comparative genome analysis.

Microbiome. 2024;12 **https://doi.org/10.1186/s40168-024-01916-8**.

Venter et al., 2004 Venter JC, Remington K, Heidelberg JF, et al.... Environmental genome shotgun sequencing of the Sargasso sea. *Science (New York, N.Y.)*. 2004;304(5667):66–74 **https://doi.org/10.1126/science.1093857**.

von Wintzingerode et al., 2000 von Wintzingerode F, Landt O, Ehrlich A. Ulf B Göbel, peptide nucleic acid-mediated pcr clamping as a useful supplement in the determination of microbial diversity. *Applied and Environmental Microbiology*. 2000;66(2):549–557 **https://doi.org/10.1128/aem.66.2.549-557.2000**.

White et al., 1990 White TJ, Bruns T, Lee S, Taylor J. *Amplification and direct sequencing of fungal ribosomal rna genes for phylogenetics* Elsevier BV 1990;315–322 **https://doi.org/10.1016/b978-0-12-372180-8.50042-1**.

Wick et al., 2022 Wick RR, Holt KE, Schneidman-Duhovny D. Polypolish: Short-read polishing of long-read bacterial genome assemblies. *PLoS Computational Biology*. 2022;18(1):e1009802 **https://doi.org/10.1371/journal.pcbi.1009802**.

Woese, 1987 Woese CR. Bacterial evolution. *Microbiological Reviews*. 1987;51(2):221–271 **https://doi.org/10.1128/mr.51.2.221-271.1987**.

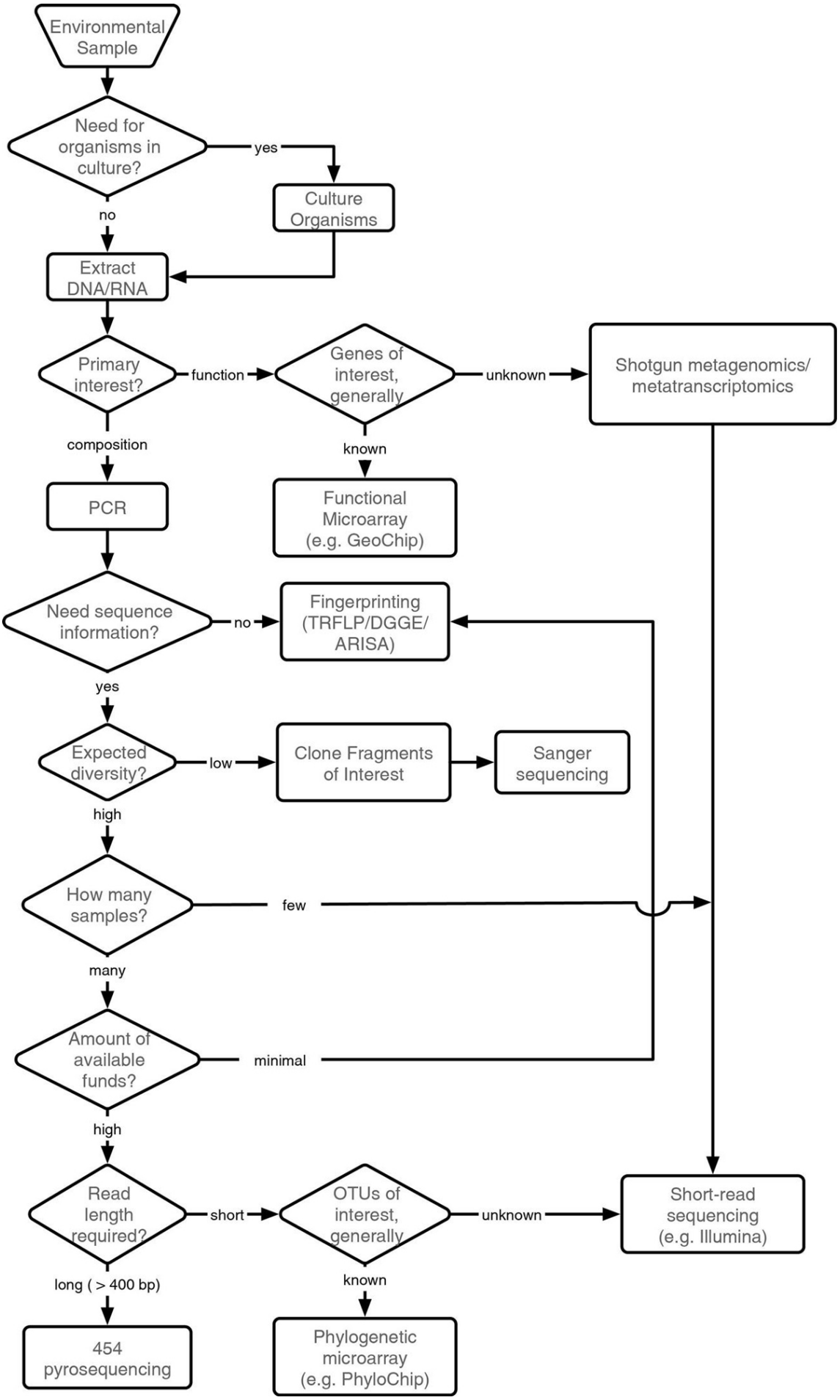
Wolfgang et al., 2013 Wolfgang WJ, Passaretti TV, Jose R, et al. Neisseria oralis sp nov., isolated from healthy gingival plaque and clinical samples. *International Journal of Systematic and Evolutionary Microbiology*. 2013;63(4):1323–1328 **https://doi.org/10.1099/ijs.0.041731-0**.

Zhang et al., 2024 Zhang Y, Li Z, Peng Y, et al. Microbiome in a ground-based analog cabin of China Space Station during a 50-

day human occupation. *ISME Communications*. 2024;4 **https://doi.org/10.1093/ismeco/ycae013**.

Zimmerman et al., 2014 Zimmerman N, Izard J, Klatt C, Zhou J, Aronson E. The unseen world: Environmental microbial sequencing and identification methods for ecologists. *Frontiers in Ecology and the Environment*. 2014;12(4):224–231 **https://doi.org/10.1890/130055**.

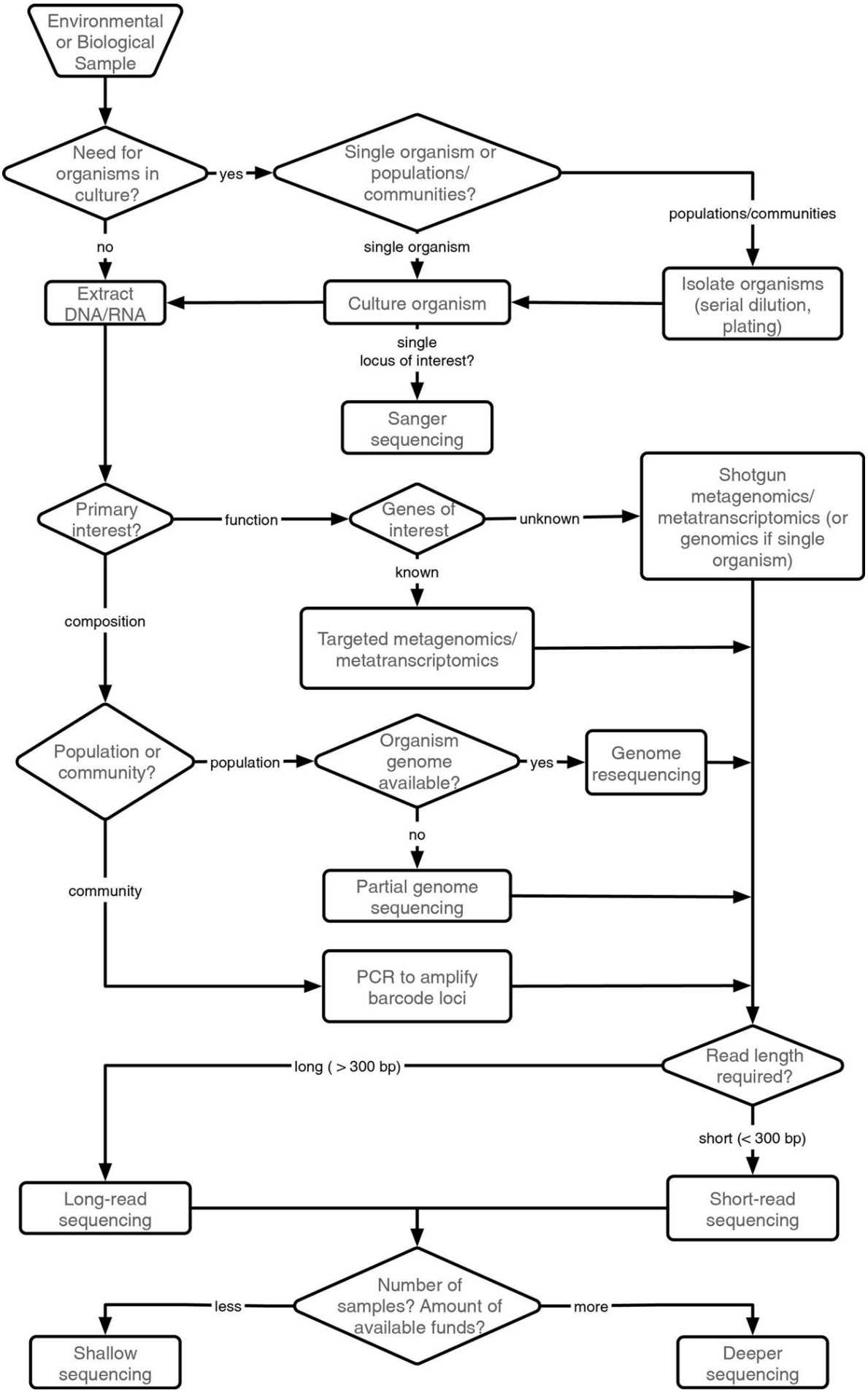
Description of Figure 2.1


Environmental Sample
Need for organisms in culture?
yes
Culture Organisms
no
Extract DNA/RNA
Primary interest?
function
Genes of interest, generally
unknown
Shotgun metagenomics/ metatranscriptomics
composition
known
PCR
Functional Microarray (e.g. GeoChip)
Need sequence information?
no
Fingerprinting (TRFLP/DGGE/ ARISA)
yes
Expected diversity?
low
Clone Fragments of Interest
Sanger sequencing
high
How many samples?
few
many
Amount of available funds?
minimal
high
Read length required?
short
OTUs of interest, generally
unknown
Short-read sequencing (e.g. Illumina)
long (> 400 bp)
known
454 pyrosequencing
Phylogenetic microarray (e.g. PhyloChip)

The flowchart starts with 'Environmental Sample' and asks 'Need for organisms in culture question mark'. If yes, it leads to 'Culture Organisms'. If no, it proceeds to 'Extract D N A/R N A'. The next decision point is 'Primary interest question mark' with options 'function' and 'composition'. 'composition' leads to 'PCR', 'Function' leads to Genes of interest generally which branches into known and unknown, 'known' leads to 'Functional Microarray (example: GeoChip)' and 'unknown' leads to 'Shotgun metagenomics or metatranscriptomics' which further leads to 'Short-read sequencing (example: Illumina)'. From 'PCR', the next decision is 'Need sequence information question mark'. If no, it leads to 'Fingerprinting (T R F L P or D G G E or ARISA)'. If yes, it proceeds to 'Expected diversity question mark'. 'Low' leads to 'Clone Fragments of Interest' and 'Sanger sequencing'. 'High' leads to 'How many samples question mark'. 'Few' leads to 'Short-read sequencing (example: Illumina)'. 'Many' leads to 'Amount of available funds question mark'. 'Minimal' leads to 'Fingerprinting (T R F L P or D G G E or ARISA)' and high lead to 'Read length required question mark' from which 'Long (greater than 400 bp)' leads to '454 pyrosequencing' and 'Short' leads to 'O T Us of interest, generally' from which further 'Unknown' leads to 'Short-read sequencing (example: Illumina)' and 'Known' leads to 'Phylogenetic microarray (example: PhyloChip)'.

Navigate back to Figure 2.1

Description of Figure 2.2


Environmental or Biological Sample
Need for organisms in culture?
yes
Single organism or populations/ communities?
populations/communities
no
single organism
Extract DNA/RNA
Culture organism
Isolate organisms (serial dilution, plating)
single locus of interest?
Sanger sequencing
Primary interest?
function
Genes of interest
unknown
Shotgun metagenomics/ metatranscriptomics (or genomics if single organism)
known
composition
Targeted metagenomics/ metatranscriptomics
Population or community?
population
Organism genome available?
yes
Genome resequencing
no
community
Partial genome sequencing
PCR to amplify barcode loci
Read length required?
long (> 300 bp)
short (< 300 bp)
Long-read sequencing
Short-read sequencing
Number of samples? Amount of available funds?
less
more
Shallow sequencing
Deeper sequencing

The flowchart begins with 'Environmental or Biological Sample' and proceeds to 'Need for organisms in culture question mark'. If yes, it leads to 'Single organism or populations or communities question mark'. For a single organism, it goes to 'Culture organism' and then 'Single locus of interest question mark'. If yes, it leads to 'Sanger sequencing'. If no, it goes to 'Extract D N A/R N A'. For populations or communities, it leads to 'Isolate organisms (serial dilution, plating)' and then 'Shotgun metagenomics/metatranscriptomics (or genomics if single organism)'. If no, it goes to 'Extract D N A or R N A' through culture organism. From 'Extract D N A or R N A', it leads to 'Primary interest question mark'. If function, it goes to 'Genes of interest'. If composition, it goes to 'Population or community question mark'. If population, it leads to 'Organism genome available question mark'. If community, it goes to 'PCR to amplify barcode loci'. From 'Genes of interest', if known, it leads to 'Targeted metagenomics or metatranscriptomics'. If unknown, it goes to 'Shotgun metagenomics or metatranscriptomics (or genomics if single organism)'. From 'Organism genome available question mark'. If yes, it goes to 'Genome resequencing'. If no, it goes to 'Partial genome sequencing' and finally to 'Read length required question mark' from both. From 'P C R to amplify barcode loci', it leads to 'Read length required question mark'. If long (greater than 300 bp), it goes to 'Read length required question mark' which further goes to 'Long-read sequencing' if long greater than 300 b p. If short (less than 300 b p), it goes to 'Short-read sequencing'. From 'Long-read sequencing' and 'Short-read sequencing', it leads to 'Number of samples question mark Amount of available funds question mark'. If less, it goes to 'Shallow sequencing'. If more, it goes to 'Deeper sequencing'.

Navigate back to Figure 2.2

CHAPTER 3

The use of metagenomics for biothreat and emerging threat applications

Dominick A. Centurioni, Alexander J. Diaz, Michael J. Perry and Christina T. Egan, New York State Department of Health, Biodefense Laboratory, Division of Infectious Diseases, Wadsworth Center, Albany, NY, United States

Abstract

Rapid detection of biothreat and emerging infectious agents is one of the core functions of a public health laboratory, as well as its ability to respond to a nefarious event. Large metagenomic next-generation sequencing (mNGS) datasets generated through deep sequencing of biothreat samples can be utilized in a variety of ways, with applications, including identification and classification of well-studied biothreat agents, detection of emerging and re-emerging infectious pathogens, and drug discovery for therapeutic use. With the increased affordability, availability, and speed of next-generation sequencing platforms over the past decade many laboratories have added mNGS capabilities to their repertoire. Additionally, with a variety of sample workflows emerging, researchers may analyze samples ranging from ancient artifacts to the air, food, and water that the public consumes each day. This chapter will review various applications of metagenomics for clinical diagnostics, detection of threats in food, and public health surveillance.

Keywords

Biodefense; biothreat agent; metagenomics; next-generation sequencing; detection of pathogens

Introduction

Rapid detection of biothreat and emerging infectious agents is one of the core functions of a public health laboratory, as well as its ability to respond to a nefarious event. Large metagenomic next-generation sequencing (mNGS) datasets generated through deep sequencing of biothreat samples can be utilized in a variety of ways, with applications including identification and classification of well-studied biothreat agents, detection of emerging and re-emerging infectious pathogens, and drug discovery for therapeutic use. Unlike next-generation sequencing (NGS), mNGS does not involve targeted sequencing allowing for a truly unbiased approach at potentially identifying all microorganisms present from a variety of specimen types. With a variety of mNGS sample workflows emerging, researchers may analyze samples ranging from ancient artifacts to the air, food, and water that the public consumes each day.

Novel or intentionally modified infectious agents may be discovered through sequencing, whereas testing using culture-based techniques or other targeted molecular methods (e.g., rtPCR) may fail to identify these pathogens. Introduction of virulence factors or antimicrobial resistance (AMR) markers through malintent could go undetected through use of targeted assays. Consequently, the US Department of Defense has indicated in the 2023 Biodefense Posture Review that there is a priority to rapidly detect and characterize emerging threats through biosurveillance and disseminate the resulting data to enable appropriate responses in a timely manner (Defense, 2023). Page 48

Clinical utility

One major application of mNGS is the detection, identification, and classification of highly infectious pathogens in clinical matrices, such as stool

and whole blood. Unlike traditional clinical diagnostic assays, unbiased mNGS can provide critical information rapidly and without limitations on the number of targets or ability to culture microbes. Researchers have made clinical diagnoses involving high-consequence pathogens such as *Bacillus anthracis* using mNGS in combination with clinical presentation, sometimes without viable organisms for standard testing and culture (Liu et al., 2023). An additional advantage is the ability to detect coinfections, ensuring that all infections are properly treated. This helps reduce disease severity and mortality that could result from undiagnosed infections (Middlebrook et al., 2022). Such coinfections have been detected in patients with tularemia through mNGS of cervical lymph node aspirates (Birdsell et al., 2018). While mNGS may not currently be appropriate for routine testing, it can be advantageous in the detection of uncommon or difficult-to-diagnose cases (Chiu & Miller, 2019). mNGS has succeeded even when diagnostic imaging and blood tests failed to identify the causative agent, *Brucella* species (Mongkolrattanothai et al., 2017). In certain cases mNGS can reduce the need for invasive procedures and offer a more cost-effective solution. For instance, scientists were able to diagnose neurobrucellosis using mNGS on cerebral spinal fluid when imaging, brain biopsies, and other tests were inconclusive (Mathias et al., 2023). Similarly, mNGS has been used to diagnose rare and unsuspected cases of ocular, hepatic, and obstetric brucellosis (Canouï et al., 2023; Lazarevic et al., 2018; Xi et al., 2022). As is the case with infections caused by many high-risk pathogens, early detection can be essential for a positive outcome. In many of the previously mentioned cases diagnosis was delayed due to the rarity of disease, the limited scope of selected assays, or the inability to culture. Additionally, high-risk procedures may be circumvented through incorporation of mNGS into routine diagnostic workflows.

mNGS data have also been used to detect biothreat agents present in unlikely clinical settings, providing the potential to inform downstream testing or public health actions. One study performed to examine the

microbiome of pooled blood donations led to the identification of pathogens, including *Coxiella burnetii*, which was confirmed by Page 49 immunological testing (Xu et al., 2018). With this information, decisions could be made to increase public health safety.

Alternatively, mNGS datasets may be used to improve treatment of infected individuals. It has been observed that infants suffering from botulism have altered gut microbiomes. Researchers have suggested the possibility of enhancing treatment through supplementation with *Bifidobacterium*, which could normalize metabolic changes resulting from colonization by *Clostridium botulinum* (Wang et al., 2023). Interestingly, mNGS of nonhuman samples may uncover natural remedies to biologic agents, such as botulinum neurotoxins (Benoit, 2018). mNGS of insect guts may even reveal uncharacterized genes which code for proteins with toxin-like structures. These toxin-producing genes can be further tested to determine toxin activity and potential clinical utility (Doxey et al., 2018).

Food and drug safety

Recent studies have examined pathogens and AMR genes present in farm settings. Researchers performing mNGS on aquafarms have identified pathogens, including *Francisella* and *Vibrio* species, and observed a greater variety in the resistome of the farm samples when compared with natural environmental samples (Seong et al., 2021). The presence of pathogens and virulence factors in such settings could impact food supply or public health and should be monitored closely. In a similar study mNGS of a pig feedlot and surrounding environmental sources was carried out to monitor the transfer of AMR genes and pathogenic bacteria. Results highlight the importance of mNGS surveillance, as human pathogens, including both *B. anthracis* and *Brucella melitensis*, were identified in both pig manure and surrounding stream water and sediment (Fang et al., 2018).

Outbreaks of highly contagious pathogens, such as highly pathogenic avian influenza, have been controlled after detection by NGS. Researchers have

suggested the use of mNGS to examine interactions between multiple attenuated live virus vaccine strains to monitor and prevent recombination and reversion to virulence (Suminda et al., 2022). A recent review of a wide variety of mNGS samples, including fecal, respiratory, blood/associated blood derivatives, skin, and organ tissue in common livestock (goats, sheep, pigs, poultry, and cattle), highlighted the ability to detect and characterize both known and novel viruses; however it was discovered that a small percentage of such Page 50 datasets are made publicly available, potentially hindering preparedness and future response efforts (Kwok et al., 2020).

Effective application of mNGS to food safety monitoring allows for detection and characterization of all known and novel pathogens without reliance on traditional culture, potentially saving time and lives in a foodborne outbreak or bioterrorism event.

A well-known challenge of the analysis of sequence data from human gut microbiome samples is filtering out the large amount of host reads sequenced. Reduction in nonpathogen reads should also be considered in food specimens. Recent comparisons of 16 S rDNA amplicon and a shotgun metagenomic approach in food samples reveal that pathogenic bacteria may be undetected when using a 16 S approach. Additionally, amplification of nucleic acid may result in a biased representation of species distribution in samples (Grützke et al., 2019). For these reasons, a metagenomic approach may be best suited for identifying pathogens in food. A recent report has documented that organisms, such as *Francisella tularensis*, can be detected in game meat; however factors such as the library preparation protocol can impact genome coverage (Grützke et al., 2019). mNGS has been used to detect and classify slow-growing or hard-to-isolate organisms, such as *Brucella* spp., allowing for a quicker response since the isolation of the organism is not required for detection. Depending on the food matrix, reducing nonpathogenic nucleic acid may be necessary to enhance sensitivity (Grützke et al., 2021). Additionally, uncommon fastidious organisms (difficult to grow and culture) may be missed using traditional methodologies requiring

culture-based enrichment with microorganism-specific selective media. mNGS could be especially useful during a foodborne outbreak to distinguish a malicious or accidental event based on the genetic characteristics of the identified contaminants.

Fieldwork

While mNGS work can be performed in a typical laboratory space, samples suspected of containing high-consequence pathogens, such as *B. anthracis*, may be best suited for workup in a secure Biosafety Level 3 (BSL-3) laboratory. Smaller sequencing devices, such as the portable Oxford Nanopore Technologies MinION, can be placed directly inside of a biosafety cabinet without the need to confirm pathogen inactivation for safe specimen transfer, thus reducing turnaround time. Page 51 Additionally, the utility of portable devices for deployment in the field could be critical to inform defense personnel, although drawbacks, such as the need for advanced bioinformatics analysis and cold reagent storage, may limit utility. Portability and ability to analyze samples in real time are essential. Using poratable nanopore sequencing devices, the US Department of Health and Human Services and the US Department of Agriculture identified select agents, such as *F. tularensis*, through mNGS during a North Atlantic Treaty Organization live agent exercise (Walter et al., 2017). A separate multicenter first-responder field exercise successfully characterized samples containing DNA from species, including *B. anthracis* and *Yersinia pseudotuberculosis* (Tyler et al., 2023). Researchers concluded that these portable devices would be valuable for use with mixed environmental samples; however distortion of species diversity was observed when comparing PCR versus PCR-free library preparations (Tyler et al., 2023). Additionally, depending on the analysis, time to results could negatively impact the potential for countermeasures (Von Fricken et al., 2023).

In another field-based mNGS study, previously weaponized encephalitic arboviruses were characterized at the strain level, further supporting the

utility of portable sequencing devices (Russell et al., 2018). In the event of a bioterrorism attack, forensic investigation through mNGS analysis of microbes may provide both detection and traceability (2013). Additionally, depending on the selected source, information including, but not limited to, cause of death and human identification, could be obtained from a single sample (Schmedes & Budowle, 2019). To monitor emerging threats at an international level, networks, such as the US Department of Defense Global Emerging Infections Surveillance program, have implemented mNGS; however it has been cautioned that continued investments into specialized workforce and infrastructure are necessary to maintain this effort and mount an effective response (Morton et al., 2023).

Surveillance

Through surveillance efforts it is possible to uncover previously unidentified hosts, reservoirs, or vectors of biothreat agents. Using a limited targeted 16 S rRNA approach researchers identified *Francisella*-like organisms in bedbugs, which had not been previously characterized as vectors of human disease (Peta et al., 2022). mNGS surveillance of environmental samples could improve our understanding of disease Page 52 ecology. For example, *F. tularensis* subsp. *holarctica*, which was previously believed to have been constrained to the northern hemisphere, was identified in possum specimens in Australia after RNA-seq analysis, which allows for deep sequencing of viral RNA to classify undiagnosed disease syndromes (Eden et al., 2017). Likewise, limited target sequencing of livestock whole blood revealed the presence of tick-borne bacterial pathogens, including *Francisella* and *Rickettsia* species, which could potentially impact livestock populations (Chaorattanakawee et al., 2022).

With the convenience of global travel it is possible that localized outbreaks can quickly threaten public health internationally. The screening of febrile travelers at large travel hubs has been proposed to quickly identify outbreaks and prevent the spread of highly infectious diseases. In one study, serum or

plasma specimens were collected from travelers with body temperatures at or above 38°C. mNGS analysis was performed side-by-side with available diagnostic testing. In addition to identifying viruses such as Ebola and dengue, mNGS identified mumps and chikungunya viruses that standard testing missed in plasma; however respiratory pathogens were not detected in these specimens, highlighting the importance of selecting the appropriate clinical matrix (e.g., serum or plasma) (Jerome et al., 2019).

Biosurveillance through air sampling in highly populated areas could be beneficial to public health outcomes, in the event of an intentional release. Atmospheric dissemination of infectious agents over long distances has been documented, originating from soil in highly cultivated environments (Rodó et al., 2014). The ability to monitor the air for high-threat pathogens potentially elucidates the causative agent of unexplained syndromes, allowing for early warning and treatment options. Aerosolized microbes in urban settings have been discovered through extraction of filters followed by mNGS, revealing temporal changes in species distribution. Through use of a genetically similar surrogate for *B. anthracis*, *Bacillus thuringiensis*, researchers concluded that an intentional release of a biothreat could be detected (Be et al., 2015). Interestingly researchers have deployed similar techniques in public settings, such as schools and athletic facilities, to detect respiratory pathogens, including influenza and coronaviruses, using the Oxford Nanopore sequencing platform (Minor et al., 2023). Detection and classification of pathogens in wastewater through rtPCR or targeted sequencing have recently been implemented in many public health laboratories; however few studies were found that utilize Page 53 unbiased mNGS to detect biothreats. Such testing may aid in detection of asymptomatic or subclinical cases and outbreaks in areas with limited access to health care (Atoui et al., 2023). Interestingly, mNGS of wastewater samples in Uganda detected or identified poxvirus among the top viral reads, highlighting the potential of mNGS for the detection of biothreat agents (O'brien et al., 2017). Additional studies utilizing an agnostic or limited targeted approach to pathogen detection in

wastewater may be an effective option for detection of biological warfare agents; however further investigation is needed.

Epidemiological investigation

A strong link exists between epidemiological investigations, surveillance, and health security. Threat-agnostic sequencing, which is a nontargeted sequencing method, has been prioritized and requires an expansion upon currently used investigation sources, such as health records (Adams et al., 2023). Recent investments have enhanced genomic sequencing capabilities, and the resulting data will be critical for epidemiological investigations. These data will enable the identification, tracking, and understanding of outbreaks as they evolve, which is vital for effective intervention strategies. Threat-agnostic sequencing capabilities can be particularly important in situations where it is not possible to isolate an agent for specialized testing. Response to the recent mpox virus outbreak in 2022 was hindered by the lack of laboratories capable of culturing positive Orthopoxvirus specimens to obtain viral isolates for sequence analysis. From the 2022 mpox outbreak, metagenomic sequencing data not only detected positive mpox specimens but also provided valuable insights, such as identifying new clades or strains, tracking mutations throughout the ongoing outbreak, and monitoring drug resistance. This additional information enhances outbreak surveillance and supports more effective public health responses and epidemiological investigations.

An epidemiological investigation recently led to the identification of rickettsial disease in an individual after traditional diagnostic testing yielded negative results. mNGS identified a small number of reads as *Rickettsia japonica* in the blood of the patient. Subsequent mNGS analysis of ticks from a dog in the patient's home led to the identification of the potential source of the *R. japonica* (Lu et al., 2024). mNGS could also supplement epidemiological investigation during foodborne outbreaks involving toxin-producing pathogens such as *C. botulinum*, especially when no

viable organism is present. Currently, isolates can be sequenced and their genomes can be compared using a SNP-based approach to infer relatedness (**Fig. 3.1**). In instances when it is not possible to obtain an axenic culture or viable organism, mNGS analysis may be essential to provide supporting evidence of relatedness. Retrospective analysis of ancient samples may also improve our understanding of disease transmission and evolution. Dental pulp has been recognized as a viable sample type for determining the cause of septicemic infections. For instance, researchers successfully identified *Yersinia pestis* Page 55 as the cause of death in human remains through mNGS of dental pulp, simultaneously revealing a connection between the depletion of a plasmid containing critical virulence factors and the decline of the second European plague (Susat et al., 2020). Another group of scientists characterized *B. melitensis* from ancient remains, shedding light on the pathogen's evolution and migration over time (Long et al., 2023). Similarly, shotgun mNGS analysis of medieval remains identified this pathogen, linking it to a specific geographic location and confirming its persistence in the region (Kay et al., 2014). Such findings could be crucial for forensic or epidemiological purposes in linking cases to a particular geographic area.

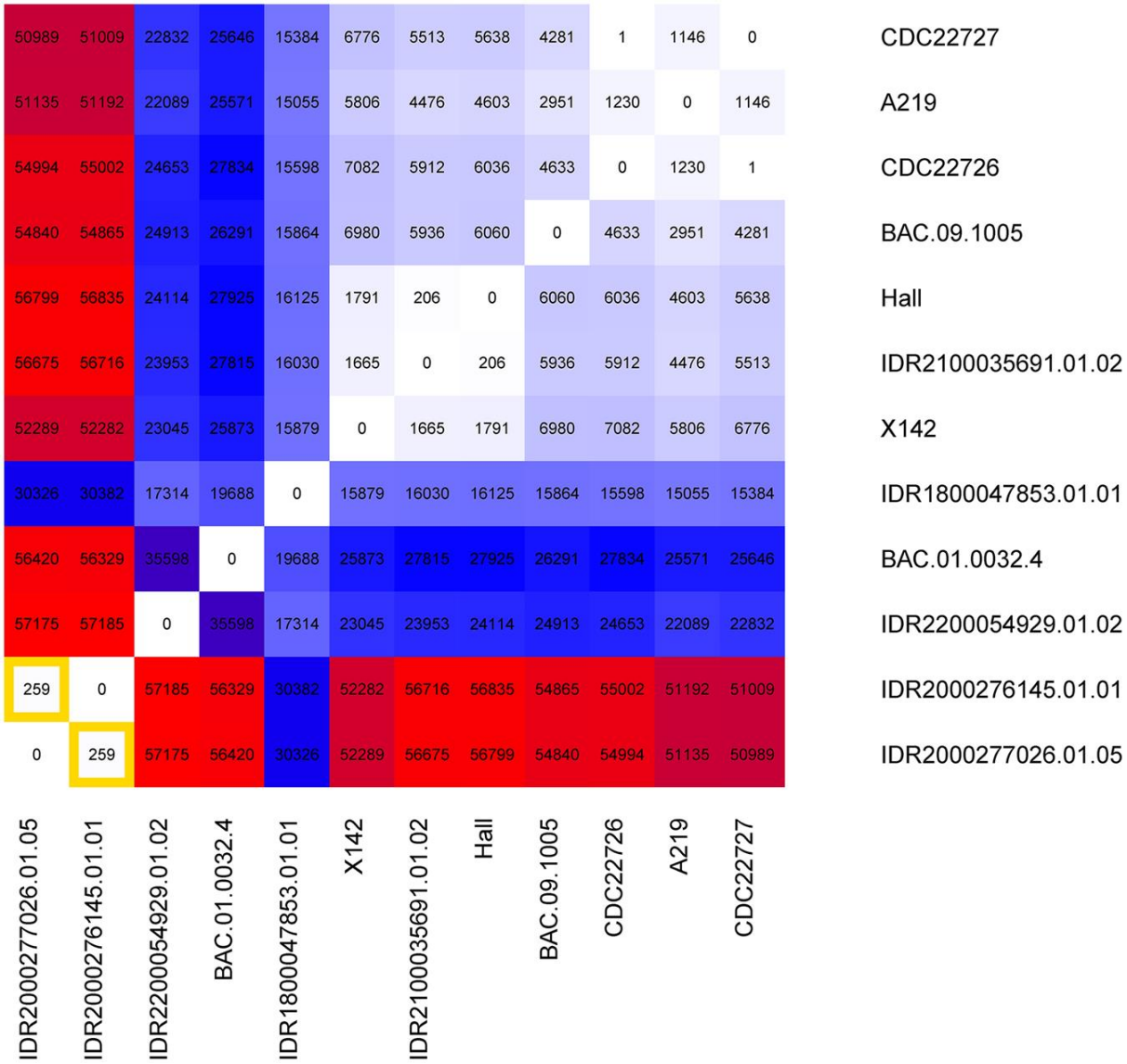


FIGURE 3.1 SNP-based heatmap constructed using the *C. botulinum* strain Hall reference (NC_009698). Labels on the *x*- and *y*-axis represent the identity of sequenced strains, while the number at the intersection of the *x*- and *y*-axis at a given location indicates the number of SNP differences identified between the two isolates. Color gradients from white to blue to red correspond to the number of SNPs identified, with white being the lowest value and red being the highest value. Two *C. botulinum* A5(B2) isolates (highlighted in yellow), IDR2000277026-01-05 and IDR2000276145-01-01, were associated with a case of foodborne botulism and were obtained from a

clinical stool and a honey sample. In comparison with other isolates in this reference grouping a low number of SNP differences may indicate relatedness. Analyses such as this may be possible without isolation of viable organism as mNGS methods improve.

Follow for extended description

Climate change and emerging threats

There has been recent concern over the potential of spillover of novel or preserved biothreat agents due to increased global temperatures. Permafrost-preserved microorganisms such as *B. anthracis* have been linked to outbreaks in wildlife (Timofeev et al., 2019). Consequently, studies have been conducted to model the potential for transmission under changing environmental conditions that may facilitate such scenarios (Stella et al., 2020). To prevent or mitigate effects associated with releases from permafrost, large-scale mNGS surveillance efforts for early detection and classification of agents are in progress (Rigou et al., 2022). While recent outbreaks have primarily occurred in Arctic environments, climate change has also been associated with severe outbreaks of *C. botulinum* in birds, driven by environmental and ecological impacts stemming from elevated temperatures in moderate climates (Gutiérrez-Arnal & Marín, 2024). Metagenomic surveillance of suitable environmental sources may be helpful to manage outbreaks and could provide an early warning for impending events.

Metagenomic next-generation sequencing approaches

mNGS methods can be sorted into two groups: amplicon and shotgun sequencing. Amplicon-based methods include tiled approaches, where overlapping amplicon segments are generated and sequenced in parallel to provide full coverage of the specific targeted region. Tiled amplicon

sequencing is the most limited, as it targets the genome of the pathogen of interest, whereas 16/18 S rDNA sequencing can examine genes across multiple organisms. Shotgun sequencing is more broad in scope and examines all sample DNA without bias (Nam et al., 2023).
Page 56 Amplicon sequencing is generally more sensitive but may not be able to detect and characterize unknown organisms, unlike shotgun sequencing (Feehery et al., 2013).

Availability of commercial kits/instruments

Unfortunately, due to the complexity of metagenomic workflows, there are currently no kits available that provide all necessary reagents and consumables for a wide range of primary specimen types to detect commensal and pathogenic bacteria using shotgun sequencing. Kits targeting the 16 S hypervariable rRNA region of the human gut microbiome can be used for targeted sequencing in stool specimens. Alternatively, enrichment of genomic regions of interest or pathogens and associated AMR markers may be targeted. Additionally, targeted amplification panels for several well-known biothreat agents exist, enhancing detection, redundancy, and species determination compared with workflows focusing solely on shotgun metagenomic sequencing (Gardner et al., 2015).

Automated instruments can take purified genetic material and perform library preparation, sequencing, and reporting to provide sequencing results. These instruments are not capable of accepting primary specimens and conducting the extraction/purification of genetic material; however they offer flexibility and efficiency.

Laboratory developed workflows

Finally, laboratories with extensive resources have the option to develop custom workflows. Publications detail the development of these workflows for targeted clinical situations, such as diagnosing patients presenting with pneumonia (He et al., 2022). mNGS workflows can also be designed for broad

applications where the etiologic agents are unknown. Development involving complex sample matrices (e.g., stool) presents challenges, such as the need to establish suitable identification criteria, which must be carefully considered. Alternatively, resource-limited laboratories may use a third-party service, to assist in workflow development.

Sample processing

The first step of mNGS wet laboratory work involves identifying and processing samples while avoiding contamination (Nam et al., 2023). Page 57 Contamination remains a significant issue in mNGS, but a method called sample-intrinsic microbial DNA found by tagging and sequencing (SIFT-seq) addresses this by using a chemical label that tags DNA from the initial sample and subsequently eliminates nontagged DNA introduced during processing, potentially mitigating contamination issues (Mzava et al., 2022).

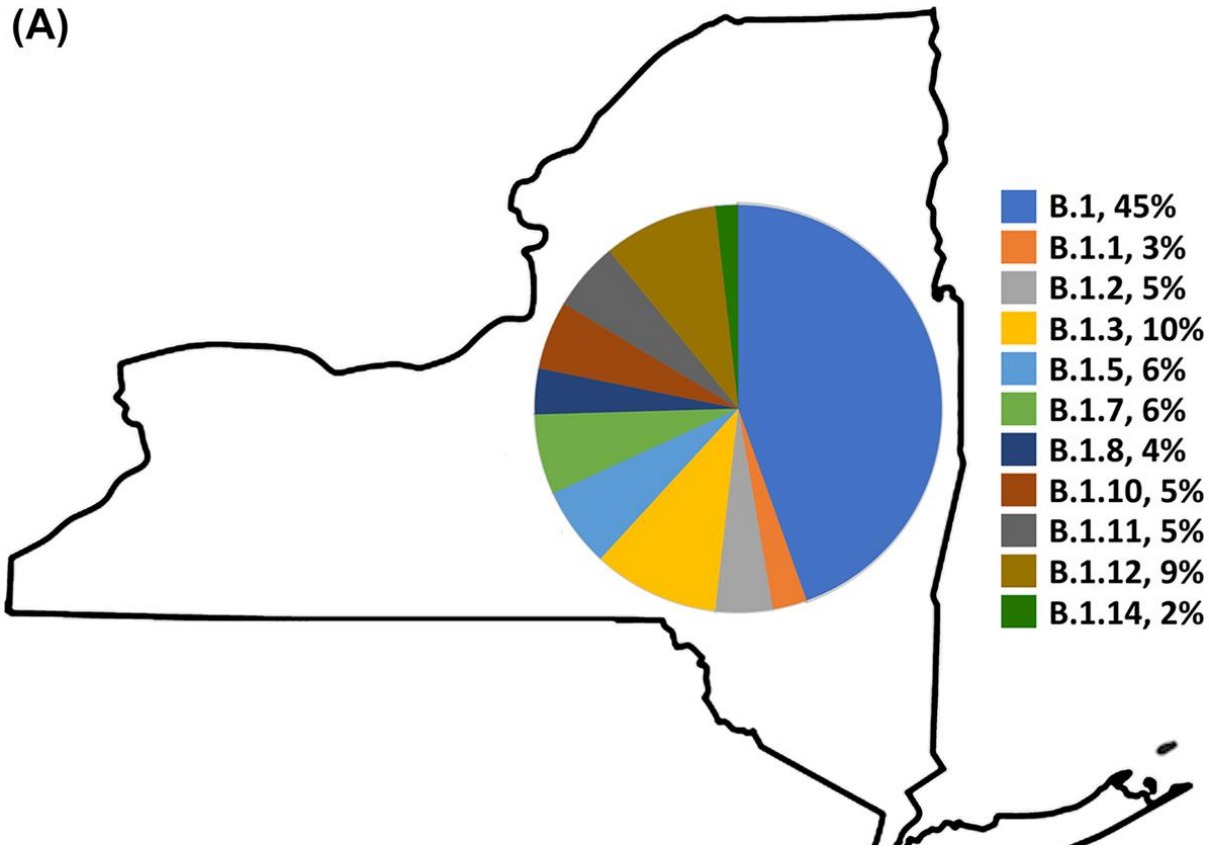
The second step of mNGS is to select and perform the appropriate extraction and manipulation of genetic material. This can vary depending on the application, but the appropriate selection and execution of this step are crucial to provide high-quality genetic material for subsequent library preparation.

The final step involves sequencing the library (Nam et al., 2023), ideally selecting a sequencing platform based on the end goal of the experiment. Traditionally, analyses that require higher accuracy and throughput often opt for short-read sequencing. In contrast applications with complex repeating genomes benefit from long-read sequencing (Adewale, 2020). Recent advancements, such as the Oxford Nanopore R10 chemistries and duplex base calling method, enhance sequencing accuracy by leveraging the complementary nature of double-stranded DNA to reinforce base calls, potentially making it a more suitable choice in various scenarios (Bogaerts et al., 2024).

Sequence analysis

Without proper analysis sequencing lacks utility. An intended purpose of sequencing analysisis is to detect and characterize pathogens of interest. This can guide downstream testing, determine sample handling (BSL-2 vs BSL-3), and influence patient treatment based on identified genetic elements.

For instance, mpox is categorized into three clades: I, IIA, and IIB (Americo et al., 2023). Clades I and IIA share a 95% nucleotide sequence identity, yet Clade I is associated with a 10% human mortality rate compared with 1% for Clade IIA (Americo et al., 2023). Furthermore, Clade IIB is less virulent than the other clades, underscoring the importance of distinguishing between clades (Americo et al., 2023). Such genetic differences between closely related pathogens can determine select agent status. Clade designation was performed on New York State specimens for these reasons, as seen below in **Fig. 3.2**. Since Clade IIB was the currently circulating strain of mpox in the United States, it was not unexpected that all mpox-positive specimens Page 58 sequenced in New York State fell within Clade IIB with little genetic variation observed.


(A)
B.1, 45%
B.1.1, 3%
B.1.2, 5%
B.1.3, 10%
B.1.5, 6%
B.1.7, 6%
B.1.8, 4%
B.1.10, 5%
B.1.11, 5%
B.1.12, 9%
B.1.14, 2%

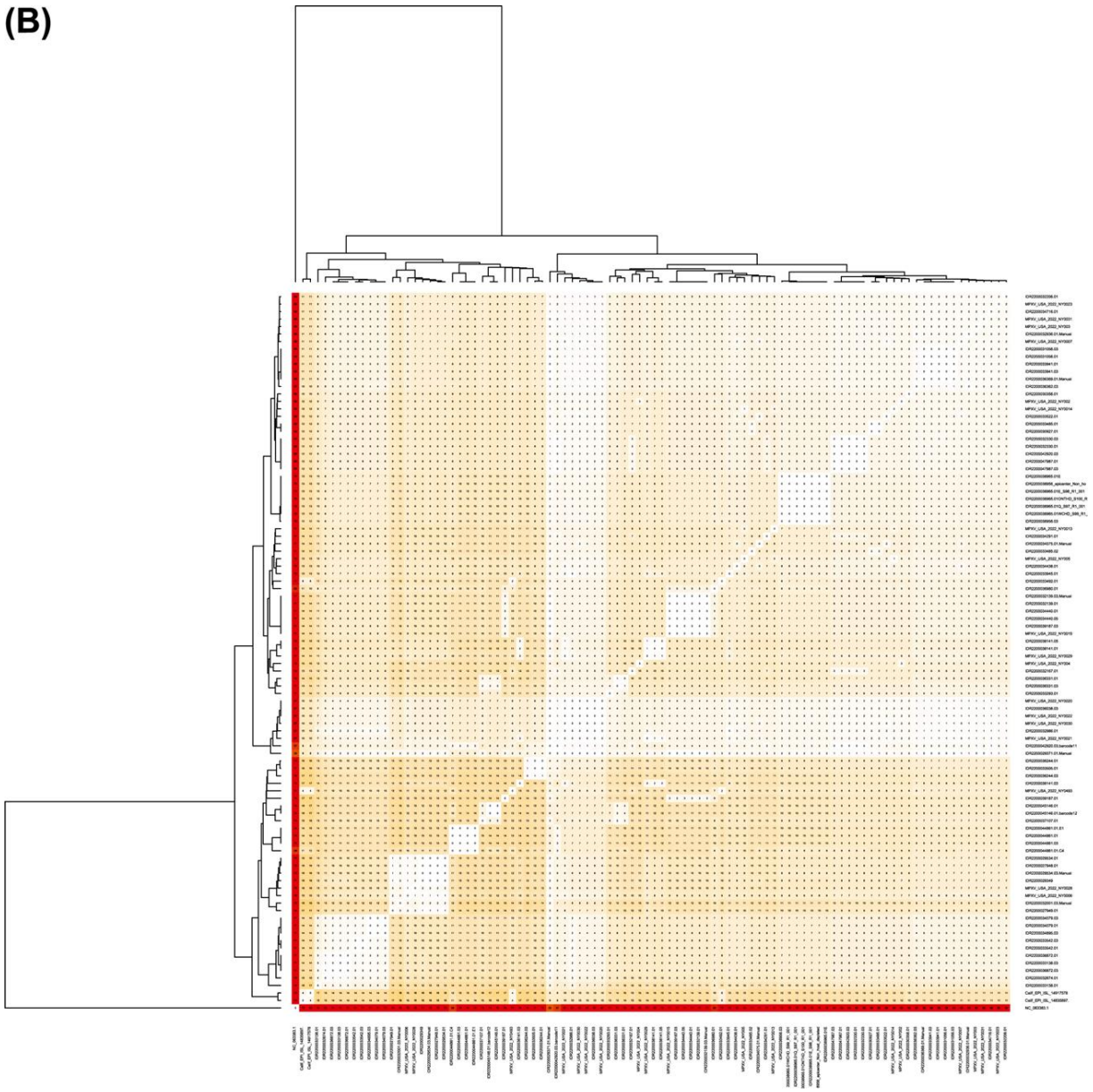


FIGURE 3.2 mNGS datasets may contain pathogen genomes with sufficient depth and coverage to allow for further characterization. As expected, little genomic variation was observed between samples sequenced during the 2022 mpox outbreak. Labels on the *x*- and *y*-axis represent the identity of sequenced strains, while the number at the intersection of the *x*- and *y*-axis at a given location indicates the number of SNP differences identified between the two samples. Color gradients from

white to yellow correspond to the number of SNPs identified, with white being the lowest value and yellow being the highest value.

Follow for extended description

Bioinformatic pipelines faciltate such analysis using open source tools including classifiers, assemblers, trimmers, and scrubbers. After data are processed a classifier categorizes reads, providing abundance, genus, and species. Care must be taken to ensure the organism of interest is present and sufficiently represented in databases to avoid misclassifications, especially between clinically significant rare organisms and genetically similar species (e.g., *Burkholderia pseudomallei* vs *Burkholderia cepacia* complex) (Tavares et al., 2020). Automated programs may flag detections of certain organisms but require verification to mitigate misinterpretations and determine clinical significance.

Two computational options support sequencing needs: local analysis using high-performance computing resources and offsite computing through cloud services. Local analysis offers immediate unmitigated access and control over sensitive data, minimizing security risks involving high-risk pathogens. Additionally, because all sensitive data Page 59 remain local, security concerns over select agent DNA sequences or sensitive medical data are limited. Unfortunately this demands substantial capital investment and technical expertise, and resource limitations can delay analyses. Cloud computing, in contrast, provides cost-effective access to extensive computational power with scalability, with a variety of interfaces. Drawbacks include complex setup and use of offsite resources necessitating de-hosting of sensitive data.

Challenges

One of the major challenges for mNGS is standardization, particularly when integrating sequencing into clinical testing. Technology, including third-

generation sequencers, is advancing at exponential rates, often Page 60 leading to tests being outdated by the time validation studies are completed. While a variety of programs/software are utilized in the analysis pipelines, there is minimal regulation of the sequence database versions and the accuracy of the pipelines.

To accurately interpret results a variety of databases must be available, depending on the specimen. Without these databases for comparison, pipeline results would not differentiate those detecting *Staphylococcus aureus* from blood (potentially pathogenic) versus skin (likely normal skin flora). Introduction of commensals during collection should also be considered. Finally, to determine the number of reads that must be present to be considered clinically significant, different thresholds and metrics must be considered, but may vary based on the matrix. Additional challenges include procurement of trained staff. Developing the necessary skills to interface with high-powered computers and associated pipelines requires significant time investiture. Data availability clauses may also clash with what is necessary for the protection of sensitive information.

Currently most laboratories utilize mNGS for research use only. When considering clinical validation, the cost per sample, limited availability of standardized analysis tools, reproducibility of results, and a lack of mNGS validation guidance can make such studies challenging. While ever-changing chemistries and consumables used with cutting-edge platforms may provide researchers with constant improvements to data quality, the time, cost, and difficulty of obtaining specimens containing biothreat agents for validation studies should be considered.

Dual use potential and cyberbiosecurity

Several ongoing surveillance efforts utilizing mNGS datasets to investigate disease acquisition and response exist. One study, BurkHostGEN, generates data from blood and water samples from the homes of melioidosis patients in endemic regions of Thailand to identify host and pathogen markers linked to

the disease (Angchagun et al., 2023). Unfortunately, while such data will be made available to improve patient outcomes, there is potential for dual use. Securing datasets containing sensitive information is critical, especially with the increased accessibility of artificial intelligence (AI) resources that can lower the barrier for an individual with malintent (Johnson et al., 2023). Using AI an individual with limited scientific knowledge could Page 61 create novel or modified harmful products affecting human or agricultural life, submit an order to a cloud laboratory, and ship the product to a target remotely. Strategies are in development to mitigate risk associated with generation and storage of sensitive datasets through use of molecular cryptography and introduction of decoy molecules to protect data before samples are sequenced (Sawaya, 2018). Additional consideration should also be given to the origin and security of the open-source analytical tools that may be installed on computing resources.

Conclusion

While mNGS is starting to become widely used for characterization of infectious agents, many laboratories that perform rapid identification of biothreat agents are not utilizing sequencing as a front-line test. There have been great strides in simplifying the technology, and the dramatic decrease in costs for mNGS will allow for broad use in a relatively short period of time. The use of mNGS will greatly improve patient care, public health action, and response and allow for the expansion and improvement of the study of pathogens and their interaction with different environments. Advancement in our knowledge of biothreat and emerging infectious agents in the next few years will support increased surveillance efforts and will be an exciting time for this field.

References

Adams et al., 2023 Adams AM, Arrazola J, Daly ER, Tompkins M. Threat agnostic epidemiology and surveillance in US Public Health Agencies: Future. *Health Security*. 2023;22:25–30 **https://doi.org/10.1089/hs.2023.0071**.

Adewale, 2020 Adewale BA. Will long-read sequencing technologies replace short-read sequencing technologies in the next 10 years?. *African Journal of Laboratory Medicine*. 2020;9(1):1340 **https://doi.org/10.1590/2318–0889202032E200032**.

Americo et al., 2023 Americo JL, Earl PL, Moss B. *Virulence differences of mpox (monkeypox) virus clades I, IIa, and IIb.1 in a small animal model*. *Proceedings of the National Academy of Sciences*. 2023;120 e2220415120 **https://doi.org/10.1073/pnas.2220415120**.

Angchagun et al., 2023 Angchagun K, Boonklang P, Chomkatekaew C, et al.... BurkHostGEN: A study protocol for evaluating variations in the Burkholderia pseudomallei and host genomes associated with melioidosis infection. *Wellcome Open Research*. 2023;8:347 **https://doi.org/10.12688/wellcomeopenres.19809.2**.

Atoui et al., 2023 Atoui A, Jourdain F, Mouly D, Cordevant C, Chesnot T, Gassilloud B. A review on mpox (monkeypox) virus shedding in wastewater and its persistence evaluation in environmental samples. *Case Studies in Chemical and Environmental Engineering*. 2023;7 **https://doi.org/10.1016/j.cscee.2023.100315**.

Be et al., 2015 Be NA, Thissen JB, Fofanov VY, et al. Metagenomic analysis of the airborne environment in urban spaces. *Microbial Ecology*. 2015;69(2):346–355 **https://doi.org/10.1007/s00248-014-0517-z**.

Benoit, 2018 Benoit RM. Botulinum neurotoxin diversity from a gene-centered view. *Toxins*. 2018;10(8):310

https://doi.org/10.3390/toxins10080310.

Bogaerts et al., 2024 Bogaerts B, Van den Bossche A, Verhaegen B, et al. Closing the gap: Oxford nanopore technologies R10 sequencing allows comparable results to Illumina sequencing for SNP-based outbreak investigation of bacterial pathogens. *Journal of Clinical Microbiology*. 2024;62(5):e01576 **https://doi.org/10.1128/jcm.01576-23**.

Birdsell et al., 2018 Birdsell DN, Özsürekci Y, Rawat A, et al. Coinfections identified from metagenomic analysis of cervical lymph nodes from tularemia patients. *BMC Infectious Diseases*. 2018;18(1):319 **https://doi.org/10.1186/s12879-018-3218-2**.

Canouï et al., 2023 Canouï E, Woerther PL, Soulier A, et al. Shotgun metagenomic screening to improve the retrospective diagnosis of undocumented intrauterine infections on embedded placenta samples: Lessons from a brucellosis case. *Infectious Diseases Now*. 2023;53 **https://doi.org/10.1016/j.idnow.2023.104770**.

Chaorattanakawee et al., 2022 Chaorattanakawee S, Wofford RN, Takhampunya R, et al. Tracking tick-borne diseases in Mongolian livestock using next generation sequencing (NGS). *Ticks and Tick-borne Diseases*. 2022;13 **https://doi.org/10.1016/j.ttbdis.2021.101845**.

Chiu and Miller, 2019 Chiu CY, Miller SA. Clinical metagenomics. *Nature Reviews Genetics*. 2019;20(6):341–355 **https://doi.org/10.1038/s41576-019-0113-7**.

Defense, 2023 Defense USDO. In: Defense DO, ed. *Biodefense Posture Review*. media.defense.gov 2023;.

Doxey et al., 2018 Doxey AC, Mansfield MJ, Montecucco C. Discovery of novel bacterial toxins by genomics and

computational biology. *Toxicon*. 2018;147:2–12 **https://doi.org/10.1016/j.toxicon.2018.02.002**.

Eden et al., 2017 Eden JS, Rose K, Ng J, et al. *Francisella* tularensis subsp holarctica in ringtail possums, Australia. *Emerging Infectious Diseases*. 2017;23(7):1198–1201 **https://doi.org/10.3201/eid2307.161863**.

Fang et al., 2018 Fang H, Han L, Zhang H, Long Z, Cai L, Yu Y. Dissemination of antibiotic resistance genes and human pathogenic bacteria from a pig feedlot to the surrounding stream and agricultural soils. *Journal of Hazardous Materials*. 2018;357:53–62 **https://doi.org/10.1016/j.jhazmat.2018.05.066**.

Feehery et al., 2013 Feehery GR, Yigit E, Oyola SO, et al. *A method for selectively enriching microbial DNA from contaminating vertebrate host DNA*. *PLoS One*. 2013;8 United States **https://doi.org/10.1371/journal.pone.0076096**.

Gardner et al., 2015 Gardner SN, Frey KG, Redden CL, et al. Targeted amplification for enhanced detection of biothreat agents by next-generation sequencing. *BMC Research Notes*. 2015;8(1):682 **https://doi.org/10.1186/s13104-015-1530-0**.

Grützke et al., 2019 Grützke J, Malorny B, Hammerl JA, et al. Fishing in the soup–pathogen detection in food safety using metabarcoding and metagenomic sequencing. *Front Microbiol*. 2019;10:1805 **https://doi.org/10.3389/fmicb.2019.01805**.

Grützke et al., 2021 Grützke J, Gwida M, Deneke C, et al. Direct identification and molecular characterization of zoonotic hazards in raw milk by metagenomics using *Brucella* as a model pathogen. *Microbial Genomics*. 2021;7 **https://doi.org/10.1099/mgen.0.000552**.

Gutiérrez-Arnal and Marín, 2024 Gutiérrez-Arnal J, Marín C. The latent threat in wild birds: *Clostridium botulinum*. *Veterinary*

Sciences. 2024;11(1):36 **https://doi.org/10.3390/vetsci11010036**.

He et al., 2022 He Y, Fang K, Shi X, et al. Enhanced DNA and RNA pathogen detection via metagenomic sequencing in patients with pneumonia. *Journal of Translational Medicine*. 2022;20(1):195 **https://doi.org/10.1186/s12967-022-03397-5**.

Jerome et al., 2019 Jerome H, Taylor C, Sreenu VB, et al. Metagenomic next-generation sequencing aids the diagnosis of viral infections in febrile returning travellers. *Journal of Infection*. 2019;79(4):383–388 **https://doi.org/10.1016/j.jinf.2019.08.003**.

Johnson et al., 2023 Johnson CD, Sinclair W, Mackelprang R. Security considerations at the intersection of engineering biology and artificial intelligence. *Engineering Biology Research Consortium* 2023; **https://doi.org/10.25498/E4J017**.

Kay et al., 2014 Kay GL, Sergeant MJ, Giuffra V, et al. Recovery of a medieval *Brucella melitensis* genome using shotgun metagenomics. *mBio*. 2014;5(4):e01337 **https://doi.org/10.1128/mBio.01337–14**.

Kwok et al., 2020 Kwok KTT, Nieuwenhuijse DF, Phan MVT, Koopmans MPG. Virus metagenomics in farm animals: A systematic review. *Viruses*. 2020;12 **https://doi.org/10.3390/v12010107**.

Lazarevic et al., 2018 Lazarevic V, Gaïa N, Girard M, et al. When bacterial culture fails, metagenomics can help: A case of chronic hepatic brucelloma assessed by next-generation sequencing. *Front Microbiol*. 2018;9:1566 **https://doi.org/10.3389/fmicb.2018.01566**.

Liu et al., 2023 Liu Y, Zheng G, Li J, et al. A case report of cutaneous Anthrax diagnosed by using a metagenomic next-generation

sequencing (mNGS) approach. *Infection and Drug Resistance*. 2023;16:3601–3606 **https://doi.org/10.2147/IDR.S411717**.

Long et al., 2023 Long GS, Hider J, Duggan AT, et al. A 14th century CE *Brucella melitensis* genome and the recent expansion of the Western Mediterranean clade. *PLoS Pathogens*. 2023;19 **https://doi.org/10.1371/journal.ppat.1011538**.

Lu et al., 2024 Lu Z, Han J, Wang Y, et al. Diagnosing and reintegrating traceability of infectious diseases via metagenomic next-generation sequencing: Study of a severe case of Rickettsia japonica infection. *Infectious Medicine*. 2024;3 **https://doi.org/10.1016/j.imj.2024.100094**.

Mathias et al., 2023 Mathias MB, Menezes FG, Fernandes GBP, et al. Neurobrucellosis mimicking primary CNS vasculitis—Should we perform CSF metagenomics before brain biopsy? A case report. *Neurology: Clinical Practice*. 2023;13 **https://doi.org/10.1212/CPJ.0000000000200167**.

Middlebrook et al., 2022 Middlebrook EA, Romero AT, Bett B, et al. Identification and distribution of pathogens coinfecting with *Brucella* spp., *Coxiella burnetii* and Rift Valley fever virus in humans, livestock and wildlife. *Zoonoses and Public Health*. 2022;69(3):175–194 **https://doi.org/10.1111/zph.12905**.

Minor et al., 2023 Minor NR, Ramuta MD, Stauss MR, et al. Metagenomic sequencing detects human respiratory and enteric viruses in air samples collected from congregate settings. *Scientific Reports*. 2023;13 **https://doi.org/10.1038/s41598-023-48352-6**.

Mongkolrattanothai et al., 2017 Mongkolrattanothai K, Naccache SN, Bender JM, et al. Neurobrucellosis: Unexpected answer from metagenomic next-generation sequencing. *Journal of the Pediatric Infectious Diseases Society*. 2017;6(4):393–398 **https://doi.org/10.1093/jpids/piw066**.

Morton et al., 2023 Morton L, Creppage K, Rahman N, et al. Challenges and opportunities in pathogen agnostic sequencing for public health surveillance: Lessons learned from the global emerging infections surveillance program. *Health Security*. 2023;22:16–24 **https://doi.org/10.1089/hs.2023.0068**.

Mzava et al., 2022 Mzava O, Cheng AP, Chang A, et al. A metagenomic DNA sequencing assay that is robust against environmental DNA contamination. *Nature Communications*. 2022;13(1):4197 **https://doi.org/10.1038/s41467-022-31654-0**.

Nam et al., 2023 Nam NN, Do HDK, Trinh KTL, Lee NY. Metagenomics: An effective approach for exploring microbial diversity and functions. *Foods*. 2023;12 **https://doi.org/10.3390/foods12112140**.

O'brien et al., 2017 O'brien E, Nakyazze J, Wu H, et al. Viral diversity and abundance in polluted waters in Kampala, Uganda. *Water Research*. 2017;127:41–49 **https://doi.org/10.1016/j.watres.2017.09.063**.

Peta et al., 2022 Peta V, Tantely LM, Potts R, Girod R, Pietri JE. A *Francisella* tularensis-like bacterium in tropical bed bugs from Madagascar. *Vector-Borne and Zoonotic Diseases*. 2022;22(1):58–61 **https://doi.org/10.1089/vbz.2021.0079**.

Rigou et al., 2022 Rigou S, Christo-Foroux E, Santini S, et al. Metagenomic survey of the microbiome of ancient Siberian permafrost and modern Kamchatkan cryosols. *microLife*. 2022;3:uqac003 **https://doi.org/10.1093/femsml/uqac003**.

Rodó et al., 2014 Rodó X, Curcoll R, Robinson M, et al. Tropospheric winds from northeastern China carry the etiologic agent of Kawasaki disease from its source to Japan. *Proceedings of the National Academy of Sciences of the United States of America*.

2014;111(22):7952–7957 **https://doi.org/10.1073/pnas.1400380111**.

Russell et al., 2018 Russell JA, Campos B, Stone J, Blosser EM, Burkett-Cadena N, Jacobs JL. Unbiased strain-typing of arbovirus directly from mosquitoes using nanopore sequencing: A field-forward biosurveillance protocol. *Scientific Reports*. 2018;8(1):5417 **https://doi.org/10.1038/s41598-018-23641-7**.

Sawaya, 2018 Sawaya S. Cryptography for genetic material. *bioRxiv* 2018;:157685 **https://doi.org/10.1101/157685**.

Schmedes and Budowle, 2019 Schmedes S, Budowle B. Microbial forensics. *Encyclopedia of Microbiology*. 2019;134 **https://doi.org/10.1016/B978-0-12-801238-3.02483-1**.

Seong et al., 2021 Seong HJ, Kim JJ, Kim T, Ahn SJ, Rho M, Sul WJ. A case study on the distribution of the environmental resistome in Korean shrimp farms. *Ecotoxicology and Environmental Safety*. 2021;227 **https://doi.org/10.1016/j.ecoenv.2021.112858**.

Stella et al., 2020 Stella E, Mari L, Gabrieli J, Barbante C, Bertuzzo E. Permafrost dynamics and the risk of anthrax transmission: A modelling study. *Scientific Reports*. 2020;10(1):16460 **https://doi.org/10.1038/s41598-020-72440-6**.

Suminda et al., 2022 Suminda GGD, Bhandari S, Won Y, et al. High-throughput sequencing technologies in the detection of livestock pathogens, diagnosis, and zoonotic surveillance. *Computational and Structural Biotechnology Journal*. 2022;20:5378–5392 **https://doi.org/10.1016/j.csbj.2022.09.028**.

Susat et al., 2020 Susat J, Bonczarowska JH, Pētersone-Gordina E, et al. Yersinia pestis strains from Latvia show depletion of the pla virulence gene at the end of the second plague pandemic.

Scientific Reports. 2020;10(1):14628 **https://doi.org/10.1038/s41598-020-71530-9**.

Tavares et al., 2020 Tavares M, Kozak M, Balola A, Sá-Correia I. Burkholderia cepacia : A feared contamination risk in water-based pharmaceutical products. *Clinical Microbiology Reviews*. 2020;33 **https://doi.org/10.1128/cmr.00139–19**.

Timofeev et al., 2019 Timofeev V, Bahtejeva I, Mironova R, et al. Insights from *Bacillus anthracis* strains isolated from permafrost in the tundra zone of Russia. *PLoS One*. 2019;14(5):e0209140 **https://doi.org/10.1371/journal.pone.0209140**.

Tyler et al., 2023 Tyler AD, McAllister J, Stapleton H, et al. Field-based detection of bacteria using nanopore sequencing: Method evaluation for biothreat detection in complex samples. *PLoS One*. 2023;18 **https://doi.org/10.1371/journal.pone.0295028**.

Von Fricken et al., 2023 Von Fricken ME, Melendrez MC, Linton YM, Takhampunya R. Editorial: Metagenomics for epidemiological surveillance in One Health. *Frontiers in Microbiology*. 2023;14 **https://doi.org/10.3389/fmicb.2023.1191946**.

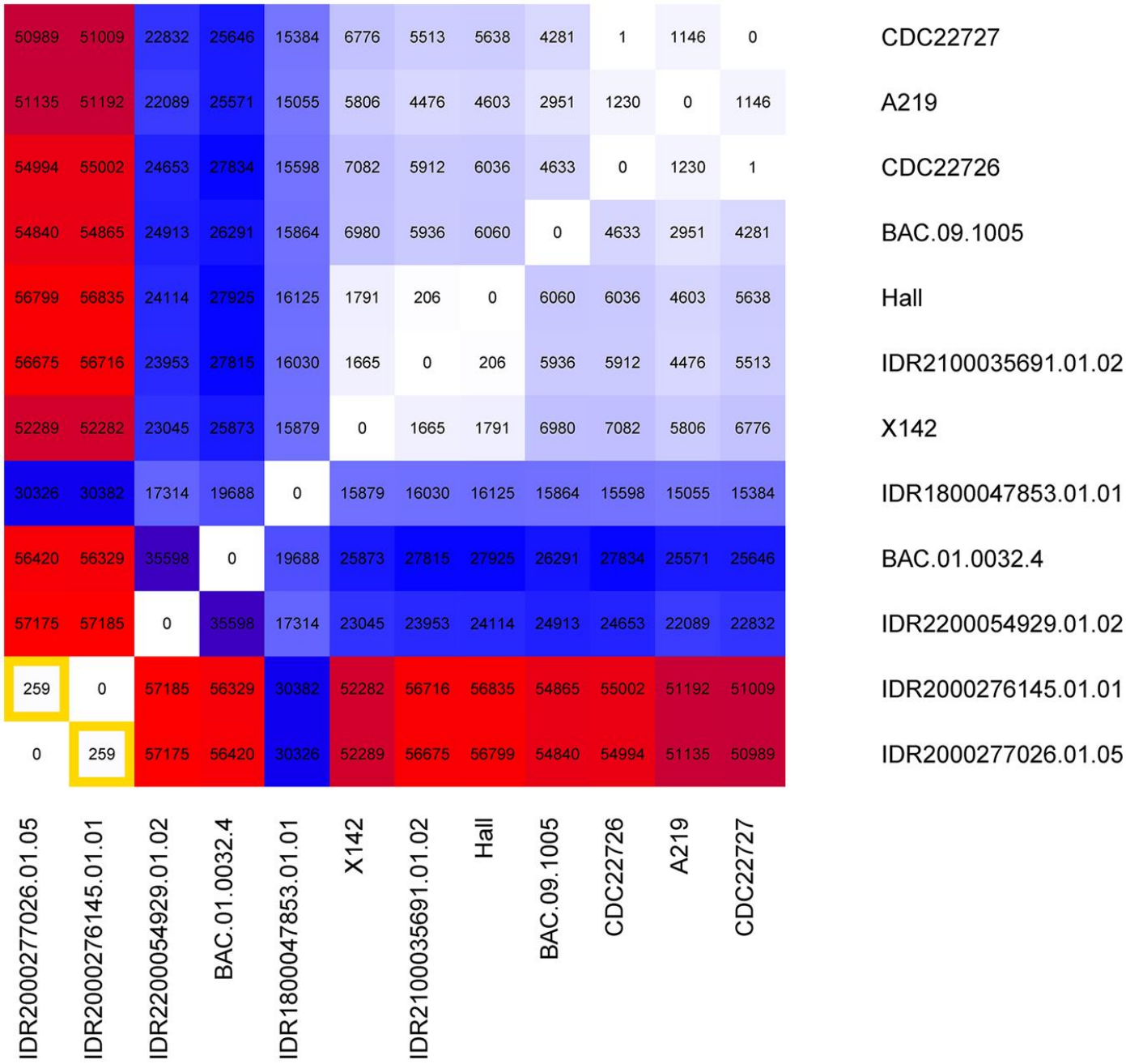
Walter et al., 2017 Walter MC, Zwirglmaier K, Vette P, et al. MinION as part of a biomedical rapidly deployable laboratory. *Journal of Biotechnology*. 2017;250:16–22 **https://doi.org/10.1016/j.jbiotec.2016.12.006**.

Wang et al., 2023 Wang D, Li K, Wang L, et al. *Dissecting and Tracing the Gut Microbiota of Infants with Botulism: A Cross Sectional and Longitudinal Study* China SSRN: SSRN; 2023; **https://doi.org/10.2139/ssrn.4508418**.

Xi et al., 2022 Xi H, Zhang L, Xu B, Liu H, Li S. Metagenomic next-generation sequencing to investigate infectious endophthalmitis of *Brucella*: A case report. *Frontiers in Medicine*. 2022;9:847143 **https://doi.org/10.3389/fmed.2022.847143**.

Xu et al., 2018 Xu M, Yang Y, Zhou Y, Liu Z, Liu Y, He M. Metagenomics in pooled plasma, with identification of potential emerging infectious pathogens. *Transfusion*. 2018;58(3):633–637 **https://doi.org/10.1111/trf.14455**.

Description of Figure 3.1



The image shows a 12 by 12 matrix of numerical values, with row identifiers on the right side. The matrix values range from 0 to 56835. Row identifiers from top to bottom: CDC22727, A219, CDC22726, BAC.09.1005, Hall, IDR2100035691.01.02, X142, IDR1800047853.01.01, BAC.01.0032.4, IDR2200054929.01.02, IDR2000276145.01.01, IDR2000277026.01.05.

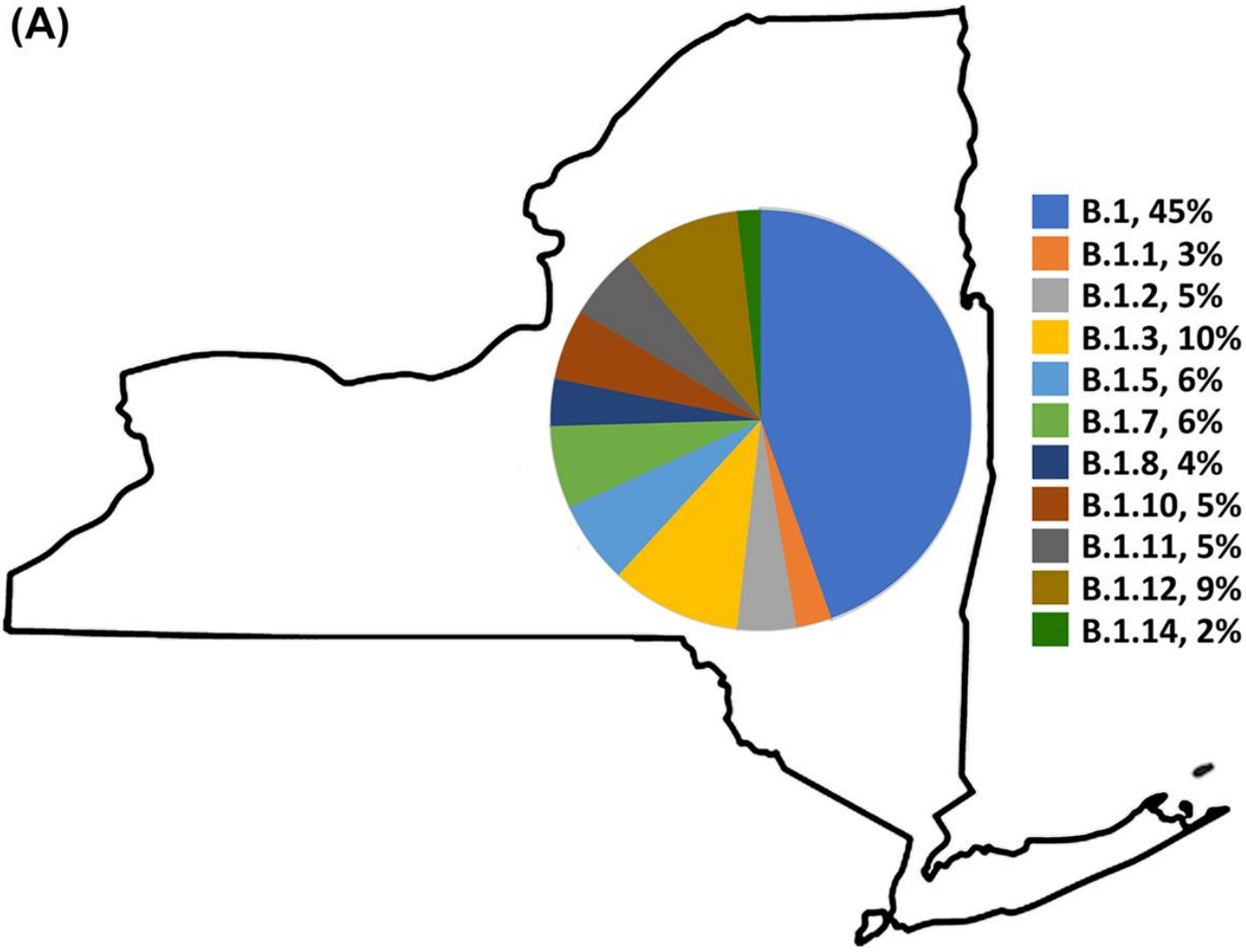
Column identifiers from left to right (shown at bottom): IDR2000277026.01.05, IDR2000276145.01.01, IDR2200054929.01.02, BAC.01.0032.4, IDR1800047853.01.01, X142, IDR2100035691.01.02, Hall, BAC.09.1005, CDC22726, A219, CDC22727. The highest values (50000 plus) are concentrated in the top-left corner and bottom of the matrix. The values generally decrease moving towards the bottom-right corner, with many cells containing smaller numbers (under 10000) or zeros. Some notable patterns: The diagonal from top-right to bottom-left contains zeros. The detailed data points are as follows: - For CDC22727, the values are: CDC22727 equals 50989, A219 equals 51135, CDC22726 equals 54994, BAC.09.1005 equals 54840, Hall equals 56799, IDR2100035691.01.02 equals 56675, X142 equals 52289, IDR1800047853.01.01 equals 50326, BAC.01.0032.4 equals 56420, IDR2200054929.01.02 equals 57175, IDR2000276145.01.01 equals 259 and IDR2000277026.01.05 equals 0. - For A219, the values are: CDC22727 equals 51009, A219 equals 51192, CDC22726 equals 55002, BAC.09.1005 equals 54865, Hall equals 56835, IDR2100035691.01.02 equals 56716, X142 equals 52282, IDR1800047853.01.01 equals 50393, BAC.01.0032.4 equals 56329, IDR2200054929.01.02 equals 57185, IDR2000276145.01.01 equals 0 and IDR2000277026.01.05 equals 259. - For CDC22726, the values are: CDC22727 equals 22832, A219 equals 22089, CDC22726 equals 24653, BAC.09.1005 equals 24913, Hall equals 24114, IDR2100035691.01.02 equals 23953, X142 equals 23045, IDR1800047853.01.01 equals 17314, BAC.01.0032.4 equals 25598, IDR2200054929.01.02 equals 0, IDR2000276145.01.01 equals 57185 and IDR2000277026.01.05 equals 57175. - For BAC.09.1005, the values are: CDC22727 equals 25646, A219 equals 25571, CDC22726 equals 27834, BAC.09.1005 equals 26291, Hall equals 27925, IDR2100035691.01.02 equals 27915, X142 equals 25873, IDR1800047853.01.01 equals 19688, BAC.01.0032.4 equals 0, IDR2200054929.01.02 equals 25598, IDR2000276145.01.01 equals 56329 and IDR2000277026.01.05 equals 56420. - For Hall, the values are: CDC22727 equals 15384, A219 equals 15055, CDC22726 equals 15598,

BAC.09.1005 equals 15864, Hall equals 16125, IDR2100035691.01.02 equals 16030, X142 equals 15879, IDR1800047853.01.01 equals 0, BAC.01.0032.4 equals 19688, IDR2200054929.01.02 equals 17314, IDR2000276145.01.01 equals 50393 and IDR2000277026.01.05 equals 50326. - For IDR2100035691.01.02, the values are: CDC22727 equals 6776, A219 equals 5806, CDC22726 equals 7082, BAC.09.1005 equals 6980, Hall equals 1791, IDR2100035691.01.02 equals 1665, X142 equals 0, IDR1800047853.01.01 equals 16030, BAC.01.0032.4 equals 27915, IDR2200054929.01.02 equals 23953, IDR2000276145.01.01 equals 56716 and IDR2000277026.01.05 equals 56675. - For X142, the values are: CDC22727 equals 5513, A219 equals 4476, CDC22726 equals 5912, BAC.09.1005 equals 5936, Hall equals 206, IDR2100035691.01.02 equals 0, X142 equals 1665, IDR1800047853.01.01 equals 16125, BAC.01.0032.4 equals 27925, IDR2200054929.01.02 equals 24114, IDR2000276145.01.01 equals 56835 and IDR2000277026.01.05 equals 56799. - For IDR1800047853.01.01, the values are: CDC22727 equals 5638, A219 equals 4603, CDC22726 equals 6036, BAC.09.1005 equals 6060, Hall equals 0, IDR2100035691.01.02 equals 206, X142 equals 1791, IDR1800047853.01.01 equals 16125, BAC.01.0032.4 equals 27925, IDR2200054929.01.02 equals 24114, IDR2000276145.01.01 equals 56835 and IDR2000277026.01.05 equals 56799. - For BAC.01.0032.4, the values are: CDC22727 equals 4281, A219 equals 2951, CDC22726 equals 4633, BAC.09.1005 equals 0, Hall equals 6060, IDR2100035691.01.02 equals 5936, X142 equals 6980, IDR1800047853.01.01 equals 15864, BAC.01.0032.4 equals 26291, IDR2200054929.01.02 equals 24913, IDR2000276145.01.01 equals 54865 and IDR2000277026.01.05 equals 54840. - For IDR2200054929.01.02, the values are: CDC22727 equals 1, A219 equals 1230, CDC22726 equals 0, BAC.09.1005 equals 4633, Hall equals 6036, IDR2100035691.01.02 equals 5912, X142 equals 7082, IDR1800047853.01.01 equals 15598, BAC.01.0032.4 equals 27834, IDR2200054929.01.02 equals 24653, IDR2000276145.01.01 equals 55002 and IDR2000277026.01.05 equals 54994. - For IDR2000276145.01.01, the values are: CDC22727 equals

1146, A219 equals 0, CDC22726 equals 1230, BAC.09.1005 equals 2951, Hall equals 4603, IDR2100035691.01.02 equals 4476, X142 equals 5806, IDR1800047853.01.01 equals 15055, BAC.01.0032.4 equals 25571, IDR2200054929.01.02 equals 22089, IDR2000276145.01.01 equals 51192 and IDR2000277026.01.05 equals 51135. - For IDR2000277026.01.05, the values are: CDC22727 equals 0, A219 equals 1146, CDC22726 equals 1, BAC.09.1005 equals 4281, Hall equals 5638, IDR2100035691.01.02 equals 5513, X142 equals 6776, IDR1800047853.01.01 equals 15384, BAC.01.0032.4 equals 25646, IDR2200054929.01.02 equals 22832, IDR2000276145.01.01 equals 51009 and IDR2000277026.01.05 equals 50989.

Navigate back to Figure 3.1

Description of Figure 3.2



The first image presents a pie chart superimposed on an outline of New York State. The chart is accompanied by a legend on the right side that provides labels and percentages for each segment. The segments of the pie chart are as follows: B.1, 45 percent B.1.1, 3 percent B.1.2, 5 percent B.1.3, 10 percent B.1.5, 6 percent B.1.7, 6 percent B.1.8, 4 percent B.1.10, 5 percent B.1.11, 5 percent B.1.12, 9 percent B.1.14, 2 percent. The second image presents a large-scale heat map visualization with an accompanying dendrogram on the left side. The heatmap is composed of a grid of small dots, where the intensity of each dot represents a value or relationship between the variables on the x and y

axes. The x-axis and y-axis both contain numerous labels, which appear to be identifiers or variables. These labels are densely packed and difficult to read individually due to their small size and quantity. The dendrogram on the left side of the heatmap shows hierarchical clustering of the y-axis variables. It branches out from left to right, grouping similar variables together. The heat map itself displays varying intensities of dots, creating patterns across the grid. Some areas show clusters of higher intensity dots, while others have more sparse or lower intensity patterns. At the left and bottom of the heat map, there are small red squares which appears like the x and y axis.

Navigate back to Figure 3.2

CHAPTER 4

Metagenomics, microbiomes, and symbioses

Maria C. Rivera, School of Life Sciences and Sustainability, Virginia Commonwealth University, Richmond, VA, United States

Abstract

It is well understood that close symbiotic interactions among prokaryotes led to the origin and evolution of the complexity of eukaryotic cells by a series of endosymbiotic events. Recent metagenomics advances applied to mammalian systems uncovered the complexity and diversity of host-associated microbiomes and their role in host health. However the extent of the host–microbiota associations among nonvertebrate eukaryotes remains understudied. This chapter discusses recent studies expanding our knowledge of the host–microbiome interactions across the tree of life and how a better understanding of these communities from an evolutionary perspective will contribute to uncovering the biological rules shaping the structures of these communities.

Keywords

Host–microbiome; microbial communities; endosymbiosis; tree of life; helminths; multiomics; symbiosis

Introduction

Metagenomic sequencing approaches have documented and highlighted the critical role of microbial communities across many habitats. Advances in

culture-independent molecular methods and next-generation sequencing (NGS) have made it possible to demonstrate the complexity and extent of the microbial communities living in and on eukaryotes. Metagenomics has revealed that the guts of most vertebrates house diverse and complex microbial communities. As a point of reference the human gut is estimated to host more than 3000 bacterial species (Human Microbiome Project Consortium, 2012). Microbial communities provide adaptive advantages to their hosts, including new metabolic capabilities, allowing the host to expand its access to environmental resources.

Many studies have highlighted the multiple benefits that the gut microbiome community provides to its host, including digestion, nutritional supplementation, vitamins, immune system modulation, and resistance to pathogens. Comparative gut microbiome studies, primarily focused on mammals, have identified diet, phylogeny, host genetics, and environment as factors that explain the observed structure and composition of the microbiome (Youngblut et al., 2019). The findings from applying omics tools to the microbial communities of the gastrointestinal (GI) tract of humans and other mammals have introduced the concepts of microbiome, microbiota, host, symbionts, and symbiosis to the broader scientific community; Douglas (2022) provides definitions of the terms.

Symbioses, the persistent associations between taxonomically different organisms, are well-established across multiple groups of organisms. These associations are recognized as drivers of evolution and innovation throughout the tree of life. Symbioses are classified based on the degree of benefits received by each member of the interacting Page 68 species: commensalism, mutualism, and parasitism. In commensalism one member of the partnership gets benefits while the other remains unaffected or only minimally affected. Mutualism occurs when both partners benefit from their association. Parasitism describes a situation in which one partner benefits at the expense of the other. The terms obligate and facultative are also used to describe the degree of dependence of the symbiont on the host; Douglas

(2022) presents a detailed description of the terminology used to classify symbioses.

Although the metagenomic sequencing era has contributed to the development of new concepts to explain the observed symbiotic interactions, it is essential to emphasize that knowledge of the important role of symbioses in biology predates this era. The term symbiosis was first defined by De Bary in 1878 as "the living together of two organisms" (Oulhen et al., 2016). By the late 1800s European scientists were already using the symbiosis concept; Archibald (2015) cites some of these early contributions. As early as 1905 Mereschkowski proposed that plastids and animal cells resulted from endosymbiotic events (Kowallik & Martin, 2021). Mereschkowski described the evolution by symbiosis as "the origin of organisms through the combination and unification of two or many beings entering into symbiosis." This is the more restrictive definition of symbiosis, in which one partner, the endosymbiont, lives inside the cell of the other partner, the host, providing access to new environments or energy sources. Endosymbiotic relationships can persist when the endosymbiont provides an evolutionary advantage, allowing the chimeric organism to better survive or conquer new ecological niches (i.e., mitochondria provide energy and plastids photosynthesis). The endosymbiont enables the host to access environments or diets that were previously inaccessible.

Recognizing that animals have established and evolved significant and extensive symbiotic interactions with microorganisms has led to new theoretical frameworks that incorporate concepts of evolution and natural selection. The extensive diversity of host–microbiome interactions uncovered during the metagenomics era has led to the proposal of the hologenome-holobiont theory of evolution (Rosenberg & Zilber-Rosenberg, 2011; Zilber-Rosenberg & Rosenberg, 2008). This theory proposes that animals, the holobiont, are a composites of host cells and the cells of the microbial symbionts, while the hologenome represents the combined genetic information of both the host and the symbionts; therefore, the holobiont

should be considered the unit of selection. Although the holobiont and hologenome concepts are widely used in Page 69 the microbiome literature, not all scientists have fully embraced the principles of the hologenome theory. Some have pointed out insufficient evidence demonstrating that natural selection acts on the host and its microbiome as a unit, which is a requirement for the system to evolve as a hologenome (Douglas & Werren, 2016).

While the hologenome-holobiont concept considers the relationship between a specific host and its persistent symbionts (Rosenberg & Zilber-Rosenberg, 2011; Zilber-Rosenberg & Rosenberg, 2008), an alternative framework suggests that what is necessary is the conservation of the collective functions of the microbial community rather than the specific set of taxa responsible for those functions (Doolittle and Inkpen, 2018). This framework posits that natural selection acts on the processes and patterns that allow for the conservation of the collective functions of the microbial communities, rather than on preservation of a particular set of taxa, as different members of the community experience varying levels of natural selection. Doolittle and Inkpen explain that the observed conservation patterns of microbial communities functions are more accurately described by differential persistence than by differential reproduction, as formulated by evolution through natural selection (Doolittle & Inkpen, 2018).

The discovery of the extensive diversity of the microbial world and its complex associations with animals has resurrected the concept of symbiosis as a core principle of contemporary biology (Gilbert, 2014). Gilbert (2014) recognized symbiosis "as the way of eukaryotic life" and proposed that "animals cannot be considered individuals by anatomical or physiological criteria because a diversity of symbionts is both present and functional in completing metabolic pathways and serving other physiological functions." The findings of the metagenomics era have influenced biologists, science historians, and philosophers of biological sciences to recognize the centrality of the symbiotic relationships in the evolution and diversification of life on

Earth, concluding that animals are symbiotic complexes of many species living together.

A contrasting perspectives on the ubiquity of microbiomes has been published by Hammer and collaborators (Hammer et al., 2019). The authors considered that there is limited information about the microbial associations in most animal species and argued that the scientific community has assumed that "all animals need a microbiome" (Hammer et al., 2019). They proposed that microbiome studies should be designed to reject the null hypothesis (i.e., the system under study lacks a microbiome) (Hammer et al., 2019). An alternative to the Page 70 generalization that all animals harbor symbionts providing essential functions to the host is the concept of the "microbial dependence continuum" (Williams & Fontaine, 2024). The proponents of the microbial dependence continuum argue that the reliance of organisms on microbiomes ranges from complete dependence to complete independence and that comparative approaches across many hosts would contribute to a better understanding of the degree of microbiome dependence by the host (Williams & Fontaine, 2024). Undoubtedly further knowledge of the prevalence of host–microbiome interdependence across the tree of life is needed to comprehend these associations' evolution, ecology, and physiology. The capacity to generate multiomics data and properly integrate these diverse data sets promises to expand the understanding of the metagenomics and microbiome fields beyond the surveys that describe the taxonomic diversity and complexity of the symbiotic communities.

Most of our understanding of symbiotic communities derives from research in mammalian systems, primarily the human gut. The diversity of mammals is relatively small; around 6400 species of mammals have been described (Ritchie, 2022), representing a small percentage of all animal species on Earth. In contrast, all insect species are estimated to be ~5.5 million (Stork, 2018). Both mammals and insects host complex symbiotic communities. A literature survey performed in September 2023 highlights the overrepresentation of mammalian gut studies in the microbiome

literature. The study quantified the number of gut microbiome publications for eight animal clades, including humans (~20,000 publications), mammals excluding humans (~1250 publications), fish and insects (~650 publications for each group), birds (~375 publications), reptiles and amphibians (<100 publications for each group), and mollusks (<50 publications) (see Figure 1 of Degregori and collaborators for more details) (Degregori et al., 2024). These findings demonstrate that our current understanding of the distribution and extent of symbioses across the eukaryotic tree of life is limited and primarily focused on mammals. All animals originates and evolved in a world dominated by prokaryotes. Understanding the diversity and composition of host-associated microbial communities across the tree of life is essential for uncovering the forces that shaped those relationships.

This chapter explores how host–microbiota associations have evolved along the endosymbiont–microbiome continuum, by focusing on the diversity and distribution of symbiotic relationships across the tree of life. It emphasizes the role of metagenomics and omics Page 71 technologies in advancing our understanding of the origin and diversification of these symbiotic interactions. By examining host-microbiota associations in invertebrates, the chapter emphasizes the importance of adopting an evolutionary perspective to gain insight into the processes that have shaped the mammalian gut microbiomes.

Endosymbioses in the origin of eukaryotes and mitochondria

Elucidating the role of symbionts in host biology requires understanding the role played by symbioses in shaping the evolution of life on Earth. Microfossil evidence estimates that the first microscopic bacteria-like organisms arose on Earth about 3.5 billion years ago (Schopf, 1993). After 3.5 billion years of evolution the prokaryotes have established diverse and abundant communities in all Earth's habitats. For the first ~2 billion years the interactions among the members of these early prokaryotic communities

allowed for establishing transient complex associations between two or more organisms. About 1.8 billion years ago, some of those interactions became permanently established, leading to the emergence of protoeukaryotes.

As one of the major transitions in the evolution of life on Earth eukaryogenesis has intrigued the scientific community for over 100 years. The best-supported hypotheses for the origin of the eukaryotes propose that eukaryotes result from multiple endosymbiotic events. Multiple phylogenetic studies have concluded that the eukaryotes are a chimeric lineage or a mosaic of genes that trace their ancestry to different prokaryotic lineages (Embley & Martin, 2006; Lake & Rivera, 1994; Rivera & Lake, 2004; Rivera et al., 1998; Williams et al., 2013). The simplest explanation for the inferred phylogenetic patterns is that the eukaryotes result from the fusion events of different prokaryotic groups. The observed chimeric gene patterns can be interpreted as the result of multiple endosymbiotic events followed by gene transfers from endosymbiont to the host genome, horizontal gene transfer, gene loss, gene duplications, gene conversions, and de novo gene inventions.

Metagenomic sequencing was instrumental in identifying the archaeal branch leading to the eukaryotic nuclear genes. Phylogenetic analysis of MAGs (metagenome-assembled genomes)-microbial genomes reconstructed from metagenomic samples from marine sediments-identified the *Lokiarchaeota* as the closest known prokaryotic relatives of the eukaryotes (Spang et al., 2015). The *Lokiarchaeota* is a lineage related to the TACK superphylum, which includes

Thaumarchaeota/*Aigarchaeota*/*Crenarchaeota*/*Korarchaeota* (Vosseberg et al., 2024). This study was instrumental in confirming the findings of culture-dependent studies that predated the metagenomic era, which indicated that eocytes or crenarchaea were the closest living relatives of the eukaryotes (Archibald, 2008; Rivera & Lake, 1992; Williams & Embley, 2014; Williams et al., 2013). The challenges imposed by the need for isolation and culturing of organisms hindered the discovery of the diversity within the

TACK superphylum. At that time, only members of the *Crenarchaeota* were isolated and cultivated (Vosseberg et al., 2024).

Reconstructing the timeline and the order of the endosymbiotic events that led to the formation of the first eukaryotic cells has been challenging because, due to the ancient nature of these events, genes have accumulated many nucleotide changes (hidden substitutions) that mask the signal of the evolutionary record. More precise phylogenomic analyses that model hidden substitutions more accurately support the hypotheses that archaea contributed to the "eukaryotic origins by founding the host lineage for the mitochondrial endosymbiont" (Williams et al., 2013). By employing a novel approach to analyzing the lengths of the branches of phylogenetic trees, Pittis & Gabaldón (2016) provided evidence supporting the late acquisition of mitochondria by a host with a genome containing a set of genes with chimeric origins, the first eukaryotic common ancestor (FECA). These studies indicate that the acquisition of mitochondria resulted from a symbiotic event that occurred after the initial prokaryotic fusion event (FECA) (Gabaldón, 2018; Pittis & Gabaldón, 2016). These endosymbiotic events led to the origin of mitochondria and the last eukaryotic common ancestor (LECA); Figure 3 of Vosseberg et al. (2024) summarizes the findings of Pittis & Gabaldón (2016) and illustrates the possible timeline of the endosymbiotic events. It is widely accepted that the origin of mitochondria occurred through the engulfment of an *Alphaproteobacteria* ancestor by the protoeukaryote, leading to the emergence of respiration in eukaryotes (Gray et al., 1999).

Endosymbioses and plastids evolution

Approximately 1.5 billion years after the origin of eukaryotes, another major endosymbiotic event occurred, leading to the origin of plastids (chloroplasts). In this event a mitochondrion-containing heterotrophic eukaryote engulfed a photosynthetic cyanobacterium, resulting in the stable integration of the endosymbiont as a photosynthetic organelle (chloroplast) and the emergence of photosynthetic eukaryotes (Keeling, 2013; Yoon et al., 2004).

The host of the endosymbiosis was the common ancestor of three extant lineages: the green and red algae and the glaucophytes. Subsequently, a fourth lineage, the land plants, emerged from the green algae lineage (Reyes-Prieto et al., 2007). These four lineages have primary plastids because their plastids originated from the primary endosymbiotic event: the engulfment of a cyanobacterium by a eukaryote (Keeling, 2013).

Later secondary and tertiary endosymbioses were established by the fusion of ancestral non-photosynthetic eukaryotes with a photosynthetic algae, resulting in the transfer of photosynthesis and chloroplasts from one micro-eukaryotic cell to another (Hwan et al., 2005; Reyes-Prieto et al., 2007). Independent secondary and tertiary endosymbioses have been identified in several lineages of unicellular eukaryotes, including dinoflagellates, diatoms, euglenids, and chlorachniophytes (Gagat et al., 2013). Plastids originated from a single eukaryotic lineage and later spread to other eukaryotic lineages by secondary endosymbioses. Thus, secondary endosymbioses occurred multiple times; however, not all eukaryotes possess plastids. The dinoflagellate *Symbiodinium* reflects the complexity of endosymbioses. While *Symbiodinium* results from an obligate relationship between diatom (algae) and cyanobacteria, they are also micro-algal endosymbionts of reef-building corals (Hehenberger et al., 2016; Liu et al., 2018).

Symbioses in protists

The vast majority of eukaryotes are microscopic, unicellular organisms collectively known as protists. Protists comprise a diverse group of phylogenetic clades, unified by the fact that they are eukaryotes not classified as animals, plants, or fungi, representing most of the eukaryotic diversity (Burki et al., 2021). Although all protists evolved after the LECA, the term does not describe a proper phylogenetic clade. The consensus from multiple phylogenomic studies has organized the eukaryotic tree of life into as many as 10 supergroups (Burki et al., 2020). The widespread use of NGS, metagenomics and metatranscriptomics has significantly increased the

sampling of protist taxa, contributing to a greater appreaciation of the extent of protist diversity. Metazoans, fungi, and land plants are members of two of the ten Page 74 supergroups, Most of the diversity within these 10 supergroups is found in protists (Burki et al., 2020, 2021).

The estimated number of described protist species ranges from 26,000 and 76,000, which is much lower than the species counts for fungi, plants, and animals. However, metabarcoding and metatranscriptomics have revealed an extensive diversity of undescribed protists (Burki et al., 2021). These estimates suggest that protists constitute the eukaryotic group with the greatest species diversity. In most ecosystems, protists play important biogeochemical and ecological roles (Burki et al., 2021). Protists are morphologically grouped into amoeboid, flagellated, and ciliated categories. They are also classified by their feeding adaptations and strategies into autotrophs, heterotrophs, or mixotrophs. Autotrophs obtain their nutrition through photosynthesis, while heterotrophs feed on other microorganisms, such bacteria. Heterotrophs can be further classified into saprotrophs, which consume decaying organic matter, and parasites.

In contrast, mixotrophs can switch between heterotrophy and autotrophy based on environmental conditions (Drew et al., 2021). Protists are abundant in soil, freshwater, and marine ecosystems. Their diverse feeding ecology and widespread distribution expose protists to numerous microbial interactions, increasing the likelihood of the evolution of permanent symbiotic associations. Despite their high diversity and abundance, relatively little information is available about their symbioses. As single-celled organisms, protists present challenges for isolation and culturing, likely contributing to the limited number of studies on their symbiotic communities.

For more than a century microscopy and morphology studies have documented prevalent symbiotic relationships among protists, bacteria, and archaea, as detailed in Görtz (2006) and Husnik and collaborators (Husnik et al., 2021). However, in the era of microbiomes and metagenomics, symbiotic associations involving protists have not received as much attention as those

involving animals. Numerous microbiome studies have included protists as members of the microbiota (e.g., microbiomes of soils and rhizosphere or the protists in the gut of termites). Soils harbor some of Earth's richest communities of protists (Burki et al., 2021).

Protists have acquired novel biochemical capabilities through prokaryotic symbiosis, including photosynthesis, nitrogen fixation, and methanogenesis. In a comprehensive review Husnik et al. (2021) summarized the findings of 240 publications that characterized the Page 75 taxonomic affiliation of the bacterial and archaeal symbionts from 298 protist hosts. These studies independently identified symbioses in almost all unicellular eukaryotic supergroups (for details of the phylogenetic diversity of the identified protist hosts and prokaryotic symbionts, see Figure 1 and Table S1 of Husnik et al. (2021)). Although the identified symbioses were more abundant in some lineages and rare in others, the observed uneven distribution may result from sampling bias, as many of the studies relied on the isolation and cultivation of the organisms (Husnik et al., 2021). Evidence of symbiotic associations is more prevalent in ciliates, amoebozoans, parabasalians, oxymonads (from termites' guts), dinoflagellates, diatoms, and trypanosomatids. With few exceptions most of the specimens studied harbor a single symbiont. Most of the symbionts identified are taxonomic classes of the phylum *Pseudomonadota*, including *Alphaproteobacteria*, *Betaproteobacteria*, *Gammaproteobacteria*, and other unclassified *Pseudomonadota*. Other less abundant taxa include *Bacteroidetes* (phylum *Bacteriodota*), *Chlamydiae* (phylum *Chlamydaeota*), *Cyanobacteria* (phylum *Cyanobacteriota*), members of domain archaea, and unknown prokaryotes. At least at the higher taxonomic levels, the diversity of the symbiotic communities of unicellular eukaryotes is similar to that of animals. The following paragraphs present some interesting examples of protist symbioses from the recent literature and highlight the contributions of metagenomics and omics tools to these discoveries.

Endosymbionts with potential for denitrification and aerobic respiration

Anaerobic protists are distributed worldwide in anoxic environments, including the guts of many animals. The mitochondria of anaerobic protists, such as ciliates, have evolved into hydrogenosomes, organelles that break down hydrogen. Methanogenic archaea endosymbionts found in anaerobic protists have developed a syntrophic, cooperative, relationship that allows the symbiont to utilize the hydrogen produced by the host hydrogenosome (Embley et al., 1995; Lewis et al., 2020). In some cases, these methanogenic symbionts are located in the cytoplasm, while in other cases they are associated with the hydrogenosomes (Husnik et al., 2021).

Recently, the discovery that endosymbionts of anaerobic ciliates can perform both anaerobic respiration through denitrification and oxygen respiration (Speth et al., 2024) has highlighted protists' extensive

metabolic diversity and potential. After microscopy imaging revealed a *Gammaproteobacteria*-like endosymbiont of an anaerobic freshwater ciliate, metagenomic sequencing identified a circular MAG (cMAG) with endosymbiotic genomic features, leading to the description of the bacterium *Candidatus Azoamicus ciliaticola* (Graf et al., 2021). Sequencing of 18 S rRNA amplicons indicated that the *Candidatus Azoamicus ciliaticola* ciliate host is an undescribed member of the class *Plagiopylea* within the phylum *Ciliophora*. Further analyses of MAGs and metatranscriptomes revealed that the endosymbiont has a highly reduced genome encoding genes for ATP generation via a respiratory denitrification pathway, suggesting these ciliates utilize the endosymbiont's generated ATP to breathe nitrate instead of oxygen (Graf et al., 2021). By using the 16 S rRNA sequences of *Candidatus Azoamicus ciliaticola* to search public metagenomic sequencing databases, four new respiratory species were identified in ~1000 datasets from globally distributed samples. The four circular MAGs (cMAGs) were assembled from metagenomic DNA isolated from groundwater samples collected in California, Ohio, and Germany. After reassembly, curation, and

annotation all four cMAGs encode genes that provide the endosymbionts with the potential for aerobic respiration and denitrification (Speth et al., 2024). The art of metagenome assembly of genomes (MAGs) is still a developing technique. Culturing the host and its endosymbionts and applying single-cell omics tools will allow for systematic testing to determine the evolutionary history and true nature of the metabolic capabilites of this interesting endosymbiotic system. If the *Candidatus Azoamicus ciliaticola* symbiotic association with a ciliate is confirmed, it will highlight the wealth of information in the metagenomics databases waiting for detailed analyses. Additionally, this case demonstrates the potential of MAGs to reveal previously unknown functions in metagenomics datasets.

Host genotype structures the microbiome of a diatom

Diatoms (supergroup *Stramenopiles*) are the world's most diverse group of algae. With a global distribution, diatoms represent a major phytoplankton group in marine and freshwater environments. They drive global biogeochemical cycles by contributing ~20% of annual Page 77 global carbon fixation (Falciatore et al., 2019). Diatoms and bacteria have coexisted for over 200 million years. Metagenomic studies of the microbiomes associated with two species and seven strains of diatoms showed each diatom species hosts a unique bacterial community, and the composition of these bacterial communities across strains of the same diatom species is highly conserved (Behringer et al., 2018). These studies have also highlighted the role of the phycosphere in shaping the bacterial community (Behringer et al., 2018; Cirri & Pohnert, 2019; Seymour et al., 2017). The phycosphere is the interface of phytoplankton–microbiome interaction analogous to the rhizosphere in plants (Seymour et al., 2017). Teasing apart the rules governing these complex microscale interactions between phytoplankton and bacteria goes beyond the metagenomic characterization of the whole

community; it requires new technologies and experimental approaches to uncover the spatial components of these interactions.

Thalassiosira rotula, a cosmopolitan marine diatom, serves as a model system for studying phytoplankton biology (Den et al., 2023). *T. rotula* exhibits high phenotypic and genotypic diversity, allowing the delineation of distinct populations. Analysis of the microbiomes across global isolates of *T. rotula* comprissing of 81 genotypes, identified a significant correlation between microbiome composition and host population. An experiment utilizing *T. rotula*, demonstrated that host genotype influences microbial community composition, suggesting that a genotypic filtering process is at play (Ahern et al., 2021). Identifying the role of the host genotype in structuring the microbiome of a diatom may imply that the factors shaping the host microbial community of human guts may have originated much earlier (Bonder et al., 2016; Goodrich et al., 2014; Rothschild et al., 2018). Because *T. rotula* is a system amenable to laboratory manipulation, microscopic imaging can potentially demonstrate the localization of the bacterial community with respect to the phycosphere and the diatom cell surface. The *T. rotula* microbiome studies highlight the benefits of investigating host–bacteria associations in their natural environment, complemented by controlled small-scale laboratory experiments to identify processes that could explain the evolution and maintenance of microbiome communities.

Choanoflagellates microbiome and the origin of metazoans

The emergence of metazoans, or animals, from unicellular protist ancestors was a major event in the evolution of life more than 600 Page 78 million years ago (Leadbeater, 2014). Morphological studies first suggested that the protist choanoflagellates are the closest living microbial ancestors of the animals, which was later confirmed by molecular phylogentic analyses (Carr et al., 2008). Choanoflagellates are free-living, flagellated bacterivorous

organisms that inhabit aquatic environments. They can exist as either single-celled organisms or as multicellular colonies (Leadbeater, 2014). They constantly interact with bacteria, and their transition to multicellularity and sexual reproduction is triggered by bacterial compounds. The proposal for the origin of metazoans suggests that the last common ancestor of the metazoans was a spherical organism with a morphology resembling the colonial form of choanoflagellates. The biology and phylogenetic position of choanoflagellates make this phylum a potential model system to study the establishment and evolution of host–microbiome interactions in a close animal relative. The microbiomes of the choanoflagellates are mainly unexplored. Whole metagenome sequencing of single-cell sorted *Bicosta minor* individuals reveled very low bacterial diversity and recovered the genome of an uncultured bacterial lineage with the genomic signatures of an obligate host-dependent bacterial associate (Needham et al., 2022). A similar methodology was used to discover the first giant virus known to infect choanoflagellates (Needham et al., 2019); viruses as members of the microbiome have not been studied extensively. A new choanoflagellate species, *Barroeca monosierra*, was isolated from Mono Lake, a hypersaline and alkaline lake in California, United States (Hake et al., 2024). The *B. monosierra* forms spherical colonies with the lumen colonized by a stable community of *Alphaproteobacteria* and *Gammaproteobacteria*. As *B. monosierra* colonies are maintained in laboratory cultures, the species has the potential to become an important model system for the study of host and bacteria interactions at the intersection of unicellularity and multicellularity (Hake et al., 2024; Woznica & Richards, 2024).

Most studies on the extent of symbiosis and microbiomes primarily focus on mammalian and vertebrate systems, with a few exceptions. However, much of eukaryotic phylogenetic diversity exists within the protists. Therefore, a more systematic exploration of the microbial communities of protits, including multiomics approaches, high-throughput culturomics, and microscopy, may contribute to a better understanding of the origins and

evolution of the mechanisms and interactions that govern host–microbe dynamics in animals. Page 79

Microbiomes and symbioses across the invertebrates

Metazoans are heterotrophic, multicellular eukaryotes. The current consensus phylogeny of the metazoans describes the relationships among 33 animal phyla (Telford et al., 2015). Each metazoan phylum is characterized by a distinct body plan and biology (Holland, 2011). The best-known animal phyla are Porifera, Cnidaria, Arthropoda, Nematoda, Annelida, Mollusca, Platyhelminthes, Echinodermata, and Chordata (which includes the vertebrates and two other phyla) (Holland, 2011). Thirty-two of the 33 extant metazoan phyla consitst of invertebrates, accounting for 97% of all animal species diversity. Estimates of vertebrate diversity are between 54,000 and 65,000 species (Wilkinson et al., 2021), representing only 3% of all animal species.

Small (<1 mm) marine planktonic and benthic (meiofauna) animals comprise most of the invertebrate phyla. The majority of animal diversity is found among eight invertebrate phyla. Based on body plan embryonic structures, and developmental patterns, the most specious invertebrate phyla are distributed across two clades: the Lophotrochozoa (Halanych et al., 1995), which includes Platyhelminthes (flatworms), Annelida (segmented worms), and Mollusca (shellfish, snails, squid, octopus), and the Ecdysozoa (Aguinaldo et al., 1997), which includes Nematoda (roundworms) and Arthropoda (insects, lobsters, crabs, shrimp, barnacles) and other smaller phyla. The Echinoderms (sea stars, urchins, sea cucumbers) and Porifera (sponges) rank among the most diverse phyla. The diversity of invertebrates' body plans, embryology, and development suggests they may have established equally diverse symbiotic associations.

Although the microbiomes of some invertebrate groups have been extensively characterized, including multiple species from most orders of

insects (Huerta-García & Álvarez-Cervantes, 2024), sponges (Griffiths et al., 2019; Kamke et al., 2013; Marulanda-Gomez et al., 2024; Moitinho-Silva et al., 2017; Thomas et al., 2016), corals (Blackall et al., 2015; Hernandez-Agreda et al., 2018; LaJeunesse et al., 2018), and *Ciona* (Urochordata, Tunicata) (Liberti et al., 2021), the microbiomes of most invertebrate phyla remain uncharacterized. Recently, the microbiomes of 1037 carefully isolated single animals from 21 invertebrate phyla were characterized using 16 S rRNA amplicon metabarcoding (Boscaro et al., 2022). The specimens were collected from various geographical locations and habitats. Like other studied Page 80 animals, they harbor a microbial community, and these communities differ from the environmental microbiomes, indicating these bacterial taxa have a preferential affinity for animal hosts. This comparative analysis of multiple specimens from 21 different marine phyla did not find evidence of a phylosymbiosis signal, which is the correlation between host phylogeny and the microbiome composition (Boscaro et al., 2022). Other smaller-scale studies have detected phylosymbiosis signals in investigations of invertebrate phyla (Dunaj et al., 2020; Leasi et al., 2024; Mazel et al., 2018; Pushpakumara et al., 2023; Qin et al., 2023; Tinker & Ottesen, 2020; van Opstal & Bordenstein, 2019). These findings suggest that more studies are needed to determine how general the phylosymbiosis pattern is across the animal tree of life.

The microbiomes of the placozoa and phoronids, two important metazoan phyla not covered in the previous study (Boscaro et al., 2022), have been characterized recently. Placozoa, microscopic marine invertebrates with the simplest animal morphology, are considered one of the early-branching metazoan phyla (Eitel et al., 2013, 2018). They lack an internal digestive system and perform external enzymatic digestion. Microscopy and culturing studies revealed the presence of intracellular symbionts in the placozoa phyla (Guidi et al., 2011). Analyses of single placozoa individuals using metagenomic, metatranscriptomic, metaproteomic, and microscopy techniques demonstrated that the placozoan *Trichoplax* sp. *H2* harbors two

intracellular symbionts (Gruber-Vodicka et al., 2019). One symbiont is an undescribed member of the family "*Candidatus* Midichloriaceae", an endosymbiotic lineage within the *Rickettsiales*. The other symbiont belongs to the *Margulisbacteria*, a phylum without cultured representatives and not known to form intracellular associations (Gruber-Vodicka et al., 2019). Previously the analysis of the sequences from the placozoa genome project suggested the presence of a rickettsial-like endosymbiont (Driscoll et al., 2013).

Phoronids (Lophotrochozoa, phylum *Phoronida*) are filter feeders, marine invertebrates that build external chitinous tubes for protection. Two recent studies are the first to explore the microbiomes from three different phoronid species collected from central California, United States and British Columbia, Canada (Ettinger & Eisen, 2024; Holt et al., 2023). Both studies utilized the sequencing of 16 S rRNA gene amplicons to characterize the composition of the bacterial microbiomes from animals, tubes, and surrounding environments (water or sediment). The studies reveled that animals and tubes harbor very similar Page 81 bacterial communities that differ from those found in the surrounding environment.

By analyzing single organisms and applying multiple omics tools, studies on the microbiomes of 21 invertebrate phyla, the placozoa, and the phoronids demonstrate how to design experiments to minimize the lack of precision in some microbiome studies (e.g., bulk metagenomic sequencing). Although microbiomes from a handful of invertebrate phyla remain uncharacterized, evidence suggests that animals from all phyla form stable symbiotic relationships with specific sets of microorganisms, that the complexity and structure of these communities vary across host species, and that these communities differ from the microbiomes of the surrounding environments.

A more comprehensive sampling of microbiomes across metazoans will provide information about the origin and distribution of factors shaping the composition and structure of microbiomes. Comparative studies have shown the influence of host-associated factors, including genetics, phylosymbiosis

and co-evolution, physiology and diet, environment, and the immune system. Understanding the prevalence of these identified factors across the tree of life will contribute to discovering general rules and processes that shape host–microbiota interactions.

Helminths: microbiomes within microbiomes

The application of metagenomics tools in parasitology has gained popularity in recent years (Wylezich et al., 2019). Although the symbiotic communities in the gastrointestinal (GI) tracts of many vertebrates are extensively characterized, little is known about the parasitic members of the vertebrate gut microbiome. Helminths, including parasitic worms that infect vertebrate GI tracts, have cohabited with their hosts' gut microbiomes for hundreds of millions of years (Rook, 2009). It is estimated that more than 100,000 species of helminths parasitize vertebrates, many of which remain undescribed or unknown, particularly among amphibians, reptiles, birds, and fish (Carlson et al., 2020). Helminths form a polyphyletic group because they include members of three invertebrate phyla: the Platyhelminthes (flatworms), the Nematoda (roundworms), and the Acanthocephala (thorny-headed worms) (Castro, 1996). The flatworms include the Trematoda (flukes) and the Cestoda (tapeworms). Helminths infect humans and livestock, causing significant health and economic impacts Page 82 (Charlier et al., 2014). The WHO estimated that in 2022,>1.5 billion people worldwide were infected by helminths (World Health Organisation, 2022). The most common helminths infecting humans are hookworms (e.g., *Necator americanus* and *Ancylostoma duodenale*), whipworms (e.g., *Trichuris trichiura*), and roundworms (e.g., *Ascaris lumbricoides*). Helminths that inhabit the gastrointestinal tract of livestock include the "barber's pole" worm *Haemonchus contortus*, the "brown stomach worm" *Teladorsagia circumcincta*, and the liver fluke *Fasciola hepatica*, as well as the equine tapeworm *Anoplocephala perfoliata*, which collectively cause devastating production losses worldwide (Charlier et al., 2014).

Parasitology is a field that has not benefited as much as others from metagenomics approaches. Metagenomics and NGS provide an alternative to traditional parasitology approaches for discovering, diagnosing, and surveilling helminth endoparasites (Antonopoulos et al., 2024; Kang et al., 2024). The Parasite Microbiome Project (PMP) (Dheilly et al., 2019) and The Global Parasite Project (Carlson et al., 2020) are two new initiatives aimed at advancing the field of parasitology by utilizing the latest omics tools. The PMP aims to uncover the complex interactions between parasites and hosts, their respective microbiota, and microbial communities in the direct environment of the parasite. The Global Parasite Project seeks to raise awareness of the need for an inventory of helminth diversity (Carlson et al., 2020). Sampling and culturing difficulties continue to pose challenges in characterizing the microbiota of endoparasites; a coordinated effort that combines classical methodology with omics advancements, is a promising strategy that will contribute to a better understanding of the role of the microbiomes of parasites. Combined strategies have demonstrated promising results in paleomicrobiology and paleoparasitology. These two fields represent novel and interesting applications of metagenomics and metabarcoding to characterize bacterial communities and determine the prevalence of helminths in ancient populations. A controlled study that integrated traditional paleoparasitogy techniques with metagenomic NGS sequencing identified roundworms, including *T. trichiura*, *Ascari* spp., and other intestinal parasites in remains from the 19th-century (Chessa et al., 2020).

Most studies on helminth-infected humans and animals focus on the impact of the parasite and the parasitic infection on the host–microbiome (Cortés et al., 2019). Investigations conducted in helminth-infected humans and animals, both under experimental and natural conditions, have identified significant associations between parasite colonization and alterations in the composition and function of the vertebrate gut microbiome. The observed alterations in the gut microbiome can affect the host immune

response and disease outcomes (Cortés et al., 2019). There is also evidence that helminth-associated bacteria may influence parasite viability and virulence (Álvarez-Izquierdo et al., 2022; Rinaldi et al., 2024). Similar to other animals, the helminths infecting the guts of vertebrates can also harbor a microbiome; however, their symbiotic microbiota has not been explored as throughly as that of their animal hosts (Brealey et al., 2022). There is a lack of understanding regarding the occurrence, structure, and functions of microbial communities inside the parasite's organs and tissues.

There is growing recognition that multicellular endoparasites of the vertebrate digestive tract harbor their own symbiotic community and are members of a host–parasite nested microbiome system (Hodži et al., 2023; Jenkins et al., 2019). Similar to other animals, associations between helminths and bacteria have been documented in both parasitic nematodes and platyhelminths. Bacterial intracellular endosymbionts in helminths have been recognized for over 40 years (Aho et al., 1991; Sironi et al., 1995; Vaughan et al., 2012). *Wolbachia* (order *Rickettsiales*, family *Anaplasmatacea*) is the most extensively studied endosymbiotic association. *Wolbachia* endosymbioses are restricted to arthropods and filarial nematodes, including the parasites *Onchocerca volvulus* (causing river blindness) and *Wuchereria bancrofti* and *Brugia malayi* (causing lymphatic filariasis) (Bouchery et al., 2013; Kaur et al., 2021; Landmann et al., 2019; Slatko et al., 2010). *Wolbachia* infects and manipulates the reproductive system of its host to facilitate the endosymbiont's propagation through generations of the parasite (vertical transmission of the endosymbiont) (Landmann et al., 2019; Werren et al., 2008).

Neorickettsia (order *Rickettsiales*, family *Anaplasmatacea*) is a genus of obligate intracellular endosymbiont found in parasitic flukes (Platyhelminthes, Digenea) that cause debilitating diseases in animals and humans (Pusterla et al., 2003; Rinaldi et al., 2024). The digeneans have complex life cycles, with vertebrates serving as the definitive host alongside several intermediate hosts, including a mollusk (Castro, 1996; Hurtrez-

Boussès et al., 2023). *Neorickettsia* is vertically transmitted through all stages of the digenean life cycle (Greiman et al., 2014). Also, it can be transmitted horizontally from the parasite to the host, Page 84 causing disease (Greiman et al., 2014). The whole genome of a *Neorickettsia* endobacterium was recovered during the assembly of the genome of the digenean *F. hepatica*, marking the first report of a *Neorickettsia* infecting cells of *F. hepatica* (McNulty et al., 2017). This unexpected discovery illustrates the challenges of identifying symbioses in parasitic worms.

Ascaris is an intestinal parasitic nematode of humans and livestock, causing more than 60,000 deaths annually in the developing world (World Health Organisation, 2022). The parasite infects the gut and interacts with the host-associated microbes. It is known that there is a "trilateral relationship between *Ascaris*, the host microbiota, and the host cells," but little is known about the *Ascaris*-associated microbiota (Midha et al., 2021). The microbiome of *Ascaris suum*, the species infecting pigs, and the microbiome of the host intestinal compartments were characterized using 16 S rRNA metabarcoding. The analysis showed that the gut of the parasite harbors a less diverse microbiome than the gut microbiome of the host (Midha et al., 2022). Similar findings were obtained from characterizing the structure of the gut microbiomes of *Ascaris lumbricoides* and its human host. The study found that the gut microbiome of adult roundworms was significantly different from that of their human hosts, both in infected and uninfected individuals. It also revealed that parasitic infections correlate with changes in the microbial structure of the host's gut (Muslim et al., 2024). Similarly, the analysis of the gut microbiome of the crater chipmunk (*Tamias amoenus cratericus*) and that of the infecting pinworms (*Nematoda, Heteroxynema cucullatum*) revealed that the pinworms possess a microbiota distinct from that of the chipmunk (Juan et al., 2025). These studies on nematodes infecting mammals suggest that the parasites have their own gut microbiota, supporting the concept of a nested microbiome system (Jenkins

et al., 2019; Midha et al., 2022). It would be interesting to confirm whether the observed pattern exists among other vertebrate hosts and parasites.

Hemonchus contortus, commonly known as the barber'spole worm, is a blood-sucking nematode parasite of the fourth stomach of ruminants and causes significant economic loss. NGS was utilized to investigate the bacterial community associated to all the stages of the parasite's life cycle and compare it with the microbiome of the host's gut (Sinnathamby et al., 2018). It was found that the microbiomes of adult worms and third-stage larvae (L3) differed from those in the fourth stomach and feces of the host. Using microscopy, bacteria were Page 85 identified in parasite eggs, the female reproductive tract, and the gut of adults and L3 larvae (Sinnathamby et al., 2018). A multiomics study aimed at understanding the effect of an *H. contortus* vaccine on the gut microbiome of the host and parasite found no differences in the microbiome profiles of parasites between vaccinated and control group animals (Palevich et al., 2023). Acanthocephalans are a group of obligate endoparasites that alternate between vertebrates and invertebrates to complete their life cycles. Similar comparative studies were not available for cestodes and acanthocephalans. The microbiomes within microbiomes add another layer of complexity to symbiotic associations: the coevolution of the vertebrate host, helminth parasites, and their respective microbiomes. Carefully designed omics experiments are crucial for understanding the occurrence, structure, and function of microbiomes within the organs and tissues of helminth parasites and the host microbiome, while also highlighting the need for more studies focusing on other parasitic groups.

Conclusions

Microbiology research over the past decades has uncovered the complexity and diversity of host-associated microbiomes, primarily in mammalian hosts; these discoveries were enabled by advancements in metagenomics technology. The large collection of metagenomic data from human subjets underscore the significant links between host health and its microbiome.

These findings raise questions about how microbiomes assemble, whether all eukaryotes posses a microbiome, and if similar patterns of microbiome assembly can be identified beyond vertebrates, particularly among the invertebrates and protists, where most of the eukaryotic species diversity exists. This chapter summarizes recent research aimed at illustrating the extent of our understanding of host–microbiota interactions among nonvertebrate eukaryotes, while discussing how an evolutionary perspective of the role of microbial symbioses in organism biology will enhance understanding the rules and processes that shape these communities. As illustrated by the studies throughout the chapter, a significant limitation of host–microbiome research across the eukaryotic tree of life is that most eukaryotic species diversity is microscopic and not amenable to culturing and isolation, which requires labor-intensive single-cell isolation techniques. Progress in host–microbiota associations will require multidisciplinary expertise to integrate the information Page 86 obtained from the application of metagenomics, metatranscriptomics, metabolomics, and metaproteomics to uncover the components and assembly rules defining these associations.

References

Aguinaldo et al., 1997 Aguinaldo AMA, Turbeville JM, Linford LS, et al. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*. 1997;387(6632):489–493 **https://doi.org/10.1038/387489a0**.

Ahern et al., 2021 Ahern OM, Whittaker KA, Williams TC, Hunt DE, Rynearson TA. Host genotype structures the microbiome of a globally dispersed marine phytoplankton. *Proceedings of the National Academy of Sciences of the United States of America*. 2021;118 **https://doi.org/10.1073/pnas.2105207118**.

Aho et al., 1991 Aho JM, Uglem GL, Moore JP, Larson OR. Research notes: Bacteria associated with the tegument of *Clinostomum marginatum* (Digenea). *Journal of Parasitology*. 1991;77(5):784–786 **https://doi.org/10.2307/3282717**.

Álvarez-Izquierdo et al., 2022 Álvarez-Izquierdo M, Fiallos E, Cociancic P, Esteban JG, Muñoz-Antoli C, Toledo R. Changes in resident microbiota associated with mice susceptibility or resistance to the intestinal trematode*Echinostoma caproni*. *Parasitology*. 2022;149(13):1781–1793 **https://doi.org/10.1017/S0031182022001366**.

Antonopoulos et al., 2024 Antonopoulos A, Gilleard JS, Charlier J. Next-generation sequencing technologies for helminth diagnostics and surveillance in ruminants: Shifting diagnostic barriers. *Trends in Parasitology*. 2024;40(6):511–526 **https://doi.org/10.1016/j.pt.2024.04.013**.

Archibald, 2008 Archibald JM. The eocyte hypothesis and the origin of eukaryotic cells. *Proceedings of the National Academy of Sciences of the United States of America*. 2008;105(51):20049–20050 **https://doi.org/10.1073/pnas.0811118106**.

Archibald, 2015 Archibald JM. Endosymbiosis and eukaryotic cell evolution. *Current Biology*. 2015;25(19):R911 **https://doi.org/10.1016/j.cub.2015.07.055**.

Behringer et al., 2018 Behringer G, Ochsenkühn MA, Fei C, Fanning J, Koester JA, Amin SA. Bacterial communities of diatoms display strong conservation across strains and time. *Frontiers in Microbiology*. 2018;9 **https://doi.org/10.3389/fmicb.2018.00659**.

Blackall et al., 2015 Blackall LL, Wilson B, Van Oppen MJH. Coral-the world's most diverse symbiotic ecosystem. *Molecular Ecology*. 2015;24(21):5330–5347 **https://doi.org/10.1111/mec.13400**.

Bonder et al., 2016 Bonder MJ, Kurilshikov A, Tigchelaar EF, et al.... The effect of host genetics on the gut microbiome. *Nature Genetics*. 2016;48(11):1407–1412 **https://doi.org/10.1038/ng.3663**.

Boscaro et al., 2022 Boscaro V, Holt CC, Van Steenkiste NWL, et al. Microbiomes of microscopic marine invertebrates do not reveal signatures of phylosymbiosis. *Nature Microbiology*. 2022;7(6):810–819 **https://doi.org/10.1038/s41564-022-01125-9**.

Bouchery et al., 2013 Bouchery T, Lefoulon E, Karadjian G, Nieguitsila A, Martin C. The symbiotic role of Wolbachia in Onchocercidae and its impact on filariasis. *Clinical Microbiology and Infection*. 2013;19(2):131–140 **https://doi.org/10.1111/1469-0691.12069**.

Brealey et al., 2022 Brealey JC, Lecaudey LA, Kodama M, et al. Microbiome “inception”: An intestinal cestode shapes a hierarchy of microbial communities nested within the host. *mBio*. 2022;13 **https://doi.org/10.1128/mbio.00679-22**.

Burki et al., 2020 Burki F, Roger AJ, Brown MW, Simpson AGB. The new tree of eukaryotes. *Trends in Ecology and Evolution*. 2020;35(1):43–55 **https://doi.org/10.1016/j.tree.2019.08.008**.

Burki et al., 2021 Burki F, Sandin MM, Jamy M. Diversity and ecology of protists revealed by metabarcoding. *Current Biology*. 2021;31(19):R1267 **https://doi.org/10.1016/j.cub.2021.07.066**.

Carlson et al., 2020 Carlson CJ, Dallas TA, Alexander LW, Phelan AL, Phillips AJ. What would it take to describe the global diversity of parasites?. *Proceedings of the Royal Society B*. 2020;287 **https://doi.org/10.1098/RSPB.2020.1841**.

Carr et al., 2008 Carr M, Leadbeater BSC, Hassan R, Nelson M, Baldauf SL. Molecular phylogeny of choanoflagellates, the sister group to Metazoa. *Proceedings of the National Academy of Sciences of the United States of America*. 2008;105(43):16641–16646 **https://doi.org/10.1073/pnas.0801667105**.

Castro, 1996 Castro G. A. (1996). Helminths: Structure, Classification, Growth, and Development. In S. Baron (Ed.), *Medical Microbiology*. (4th ed.). University of Texas Medical Branch at Galveston.

Charlier et al., 2014 Charlier J, van der Voort M, Kenyon F, Skuce P, Vercruysse J. Chasing helminths and their economic impact on farmed ruminants. *Trends in Parasitology*. 2014;30(7):361–367 **https://doi.org/10.1016/j.pt.2014.04.009**.

Chessa et al., 2020 Chessa D, Murgia M, Sias E, et al. Metagenomics and microscope revealed *T. trichiura* and other intestinal parasites in a cesspit of an Italian nineteenth century aristocratic palace. *Scientific Reports*. 2020;10 **https://doi.org/10.1038/s41598-020-69497-8**.

Cirri and Pohnert, 2019 Cirri E, Pohnert G. Algae–bacteria interactions that balance the planktonic microbiome. *New Phytologist*. 2019;223(1):100–106 **https://doi.org/10.1111/nph.15765**.

Cortés et al., 2019 Cortés A, Peachey L, Scotti R, Jenkins TP, Cantacessi C. Helminth-microbiota cross-talk—A journey through the vertebrate digestive system. *Molecular and Biochemical Parasitology*. 2019;233 **https://doi.org/10.1016/J.MOLBIOPARA.2019.111222**.

Degregori et al., 2024 Degregori S, Wang X, Kommala A, et al. Comparative gut microbiome research through the lens of ecology: Theoretical considerations and best practices. *Biological Reviews* 2024; **https://doi.org/10.1111/brv.13161**.

Den et al., 2023 Den TQ, Neu TR, Sultana S, Giebel HA, Simon M, Billerbeck S. Distinct glycoconjugate cell surface structures make the pelagic diatom Thalassiosira rotula an attractive habitat for bacteria. *Journal of Phycology*. 2023;59(2):309–322 **https://doi.org/10.1111/jpy.13308**.

Dheilly et al., 2019 Dheilly NM, Martínez JM, Rosario K, et al. Parasite microbiome project: Grand challenges. *PLoS Pathogens*. 2019;15 **https://doi.org/10.1371/journal.ppat.1008028**.

Doolittle and Inkpen, 2018 Doolittle WF, Inkpen SA. Processes and patterns of interaction as units of selection: An introduction to ITSNTS thinking. *Proceedings of the National Academy of Sciences*. 2018;115(16):4006–4014 **https://doi.org/10.1073/pnas.1722232115**.

Douglas, 2022 Douglas AE. The Diversity of Insect-Microbial Associations. *Walter de Gruyter GmbH* 2022;:23–68 **https://doi.org/10.1515/9780691236230-003**.

Douglas and Werren, 2016 Douglas AE, Werren JH. Holes in the hologenome: Why host-microbe symbioses are not holobionts. *mBio*. 2016;7 **https://doi.org/10.1128/mBio.02099-15**.

Drew et al., 2021 Drew GC, Stevens EJ, King KC. Microbial evolution and transitions along the parasite–mutualist continuum. *Nature Reviews Microbiology*. 2021;19(10):623–638 **https://doi.org/10.1038/s41579-021-00550-7**.

Driscoll et al., 2013 Driscoll T, Gillespie JJ, Nordberg EK, Azad AF, Sobral BW. Bacterial DNA sifted from the *Trichoplax adhaerens* (Animalia: Placozoa) genome project reveals a putative rickettsial endosymbiont. *Genome Biology and Evolution*. 2013;5(4):621–645 **https://doi.org/10.1093/gbe/evt036**.

Dunaj et al., 2020 Dunaj SJ, Bettencourt BR, Garb JE, Brucker RM. Spider phylosymbiosis: Divergence of widow spider species and

their tissues' microbiomes. *BMC Evolutionary Biology*. 2020;20 **https://doi.org/10.1186/s12862-020-01664-x**.

Eitel et al., 2013 Eitel M, Osigus HJ, DeSalle R, Schierwater B. Global diversity of the placozoa. *PLoS One*. 2013;8 **https://doi.org/10.1371/journal.pone.0057131**.

Eitel et al., 2018 Eitel M, Francis WR, Varoqueaux F, et al. Comparative genomics and the nature of placozoan species. *PLoS Biology*. 2018;16(7):e2005359 **https://doi.org/10.1371/JOURNAL.PBIO.2005359**.

Embley et al., 1995 Embley TM, Finlay BJ, Dyal PL, Hirt RP, Wilkinson M, Williams AG. Multiple origins of anaerobic ciliates with hydrogenosomes within the radiation of aerobic ciliates. *Proceedings of the Royal Society B: Biological Sciences*. 1995;262(1363):87–93 **https://doi.org/10.1098/rspb.1995.0180**.

Embley and Martin, 2006 Embley TM, Martin W. Eukaryotic evolution, changes and challenges. *Nature*. 2006;440(7084):623–630 **https://doi.org/10.1038/nature04546**.

Ettinger and Eisen, 2024 Ettinger CL, Eisen JA. *Phoronids and their tubes harbor distinct microbiomes compared to surrounding sediment* bioRxiv, United States 2024; **https://doi.org/10.1101/2024.05.28.596327**.

Falciatore et al., 2019 Falciatore A, Jaubert M, Bouly JP, Bailleul B, Mock T. Diatom molecular research comes of age: Model species for studying phytoplankton biology and diversity. *The Plant Cell*. 2019;32(3):547 **https://doi.org/10.1105/TPC.19.00158**.

Gabaldón, 2018 Gabaldón T. Relative timing of mitochondrial endosymbiosis and the “pre-mitochondrial symbioses” hypothesis. *IUBMB Life*. 2018;70(12):1188–1196 **https://doi.org/10.1002/iub.1950**.

Gagat et al., 2013 Gagat P, Bodył A, MacKiewicz P, Stiller JW. *Tertiary plastid endosymbioses in dinoflagellates*. *Endosymbiosis* Poland: Springer-Verlag Wien; 2013;233–290 **http://doi.org/10.1007/978-3-7091-1303-5**.

Gilbert, 2014 Gilbert SF. Symbiosis as the way of eukaryotic life: The dependent co-origination of the body. *Journal of Biosciences*. 2014;39(2):201–209 **https://doi.org/10.1007/s12038-013-9343-6**.

Goodrich et al., 2014 Goodrich JK, Waters JL, Poole AC, et al. Human genetics shape the gut microbiome. *Cell*. 2014;159(4):789–799 **https://doi.org/10.1016/j.cell.2014.09.053**.

Görtz, 2006 Görtz H-D. *Symbiotic Associations Between Ciliates and Prokaryotes* Springer Science and Business Media LLC 2006;364–402 **https://doi.org/10.1007/0-387-30741-9_15**.

Graf et al., 2021 Graf JS, Schorn S, Kitzinger K, et al. Anaerobic endosymbiont generates energy for ciliate host by denitrification. *Nature*. 2021;591(7850):445–450 **https://doi.org/10.1038/s41586-021-03297-6**.

Gray et al., 1999 Gray MW, Burger G, Lang BF. Mitochondrial evolution. *Science (New York, N.Y.)*. 1999;283(5407):1476–1481 **https://doi.org/10.1126/science.283.5407.1476**.

Greiman et al., 2014 Greiman SE, Tkach VV, Pulis E, Fayton TJ, Curran SS. Large scale screening of digeneans for *Neorickettsia* endosymbionts using real-time PCR reveals new *Neorickettsia* genotypes, host associations and geographic records. *PLoS One*. 2014;9 **https://doi.org/10.1371/journal.pone.0098453**.

Griffiths et al., 2019 Griffiths SM, Antwis RE, Lenzi L, et al. Host genetics and geography influence microbiome composition in the sponge Ircinia campana. *Journal of Animal Ecology*.

2019;88(11):1684–1695 **https://doi.org/10.1111/1365-2656.13065**.

Gruber-Vodicka et al., 2019 Gruber-Vodicka HR, Leisch N, Kleiner M, et al. Two intracellular and cell type-specific bacterial symbionts in the placozoan Trichoplax H2. *Nature Microbiology*. 2019;4(9):1465–1474 **https://doi.org/10.1038/s41564-019-0475-9**.

Guidi et al., 2011 Guidi L, Eitel M, Cesarini E, Schierwater B, Balsamo M. Ultrastructural analyses support different morphological lineages in the phylum placozoa Grell, 1971. *Journal of Morphology*. 2011;272(3):371–378 **https://doi.org/10.1002/jmor.10922**.

Hake et al., 2024 Hake KH, West PT, McDonald K, et al. A large colonial choanoflagellate from Mono Lake harbors live bacteria. *mBio*. 2024;15 **https://doi.org/10.1128/mbio.01623-24**.

Halanych et al., 1995 Halanych KM, Bacheller JD, Aguinaldo AMA, Liva SM, Hillis DM, Lake JA. Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. *Science (New York, N.Y.)*. 1995;267(5204):1641–1643 **https://doi.org/10.1126/science.7886451**.

Hammer et al., 2019 Hammer TJ, Sanders JG, Fierer N. Not all animals need a microbiome. *FEMS Microbiology Letters*. 2019;366 **https://doi.org/10.1093/femsle/fnz117**.

Hehenberger et al., 2016 Hehenberger E, Burki F, Kolisko M, Keeling PJ. Functional relationship between a dinoflagellate host and its diatom endosymbiont. *Molecular Biology and Evolution*. 2016;33(9):2376–2390 **https://doi.org/10.1093/molbev/msw109**.

Hernandez-Agreda et al., 2018 Hernandez-Agreda A, Leggat W, Bongaerts P, Herrera C, Ainsworth TD. Rethinking the coral microbiome: Simplicity exists within a diverse microbial

biosphere. *mBio*. 2018;9 **https://doi.org/10.1128/mBio.00812-18**.

Hodži et al., 2023 Hodži A, Dheilly NM, Cabezas-Cruz A, Berry D. The helminth holobiont: A multidimensional host–parasite–microbiota interaction. *Trends in Parasitology*. 2023;39(2):91–100 **https://doi.org/10.1016/j.pt.2022.11.012**.

Holland, 2011 Holland P. *The Animal Kingdom: A Very Short Introduction* Oxford University Press 2011; **https://doi.org/10.1093/actrade/9780199593217.001.0001**.

Holt et al., 2023 Holt CC, Dhaliwal S, Na I, Mtawali M, Boscaro V, Keeling P. Spatial compartmentalisation of bacteria in phoronid microbiomes. *Nature Research, Canada Scientific Reports*. 2023;13 **https://doi.org/10.1038/s41598-023-45652-9**.

Huerta-García and Álvarez-Cervantes, 2024 Huerta-García A, Álvarez-Cervantes J. The gut microbiota of insects: A potential source of bacteria and metabolites. *International Journal of Tropical Insect Science*. 2024;44(1):13–30 **https://doi.org/10.1007/s42690-023-01147-8**.

Human Microbiome Project Consortium, 2012 Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*. 2012;486(7402):207–214 **https://doi.org/10.1038/nature11234**.

Hurtrez-Boussès et al., 2023 Hurtrez-Boussès S, Alba A, Alda P, et al. *Overview of Interactions Between Parasitic Digenea and Their Molluscan Hosts, with Special Emphasis on the Lymnaeidae* Springer Science and Business Media LLC 2023;331–357 **https://doi.org/10.1007/978-3-031-30292-3_12**.

Husnik et al., 2021 Husnik F, Tashyreva D, Boscaro V, George EE, Lukeš J, Keeling PJ. Bacterial and archaeal symbioses with protists. *Current Biology*. 2021;31(13):R862 **https://doi.org/10.1016/j.cub.2021.05.049**.

Hwan et al., 2005 Hwan SY, Hackett JD, Van Dolah FM, Nosenko T, Lidie KL, Bhattacharya D. Tertiary endosymbiosis driven genome evolution in dinoflagellate algae. *Molecular Biology and Evolution*. 2005;22(5):1299–1308 **https://doi.org/10.1093/molbev/msi118**.

Jenkins et al., 2019 Jenkins TP, Brindley PJ, Gasser RB, Cantacessi C. Helminth microbiomes—A hidden treasure trove?. *Trends in Parasitology*. 2019;35(1):13–22 **https://doi.org/10.1016/j.pt.2018.10.007**.

Juan et al., 2025 Juan PAS, Palma L, Bell KC. Pinworm microbiomes are distinct from their chipmunk host gut microbiota. *Journal of Mammalogy* 2025;:1–10 **https://doi.org/10.1093/jmammal/gyae156**.

Kamke et al., 2013 Kamke J, Sczyrba A, Ivanova N, et al. Single-cell genomics reveals complex carbohydrate degradation patterns in poribacterial symbionts of marine sponges. *The ISME Journal*. 2013;7(12):2287–2300 **https://doi.org/10.1038/ismej.2013.111**.

Kang et al., 2024 Kang D, Choi JH, Kim M, et al. Optimization of 18S rRNA metabarcoding for the simultaneous diagnosis of intestinal parasites. *Scientific Reports*. 2024;14 **https://doi.org/10.1038/s41598-024-76304-1**.

Kaur et al., 2021 Kaur R, Shropshire JD, Cross KL, et al. Living in the endosymbiotic world of Wolbachia: A centennial review. *Cell Host and Microbe*. 2021;29(6):879–893 **https://doi.org/10.1016/j.chom.2021.03.006**.

Keeling, 2013 Keeling PJ. The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. *Annual Review of Plant Biology*. 2013;64:583–607 **https://doi.org/10.1146/annurev-arplant-050312-120144**.

Kowallik and Martin, 2021 Kowallik KV, Martin WF. The origin of symbiogenesis: An annotated English translation of Mereschkowky's 1910 paper on the theory of two plasma lineages. *Bio Systems*. 2021;199 **https://doi.org/10.1016/j.biosystems.2020.104281**.

LaJeunesse et al., 2018 LaJeunesse TC, Parkinson JE, Gabrielson PW, et al. Systematic revision of Symbiodiniaceae highlights the antiquity and diversity of coral endosymbionts. *Current Biology*. 2018;28(16):2570 **https://doi.org/10.1016/j.cub.2018.07.008**.

Lake and Rivera, 1994 Lake JA, Rivera MC. Was the nucleus the first endosymbiont?. *Proceedings of the National Academy of Sciences of the United States of America*. 1994;91(8):2880–2881 **https://doi.org/10.1073/pnas.91.8.2880**.

Landmann et al., 2019 Landmann F, Cossart P, Roy CR, Sansonetti P. The Wolbachia endosymbionts. *Microbiology Spectrum*. 2019;7 **https://doi.org/10.1128/microbiolspec.bai-0018-2019**.

Leadbeater, 2014 Leadbeater BSC. *The Choanoflagellates: Evolution, Biology and Ecology* Cambridge University Press 2014; **https://doi.org/10.1017/CBO9781139051125**.

Leasi et al., 2024 Leasi F, Eckert EM, Norenburg JL, et al. Microbiota associated with ototyphlonemertes species (nemertea, hoplonemertea, monostilifera, ototyphlonemertidae) reveal evidence of phylosymbiosis. *Ecology and Evolution*. 2024;14 **https://doi.org/10.1002/ece3.70471**.

Lewis et al., 2020 Lewis WH, Lind AE, Sendra KM, et al. Convergent evolution of hydrogenosomes from mitochondria by gene

transfer and loss. *Molecular Biology and Evolution*. 2020;37(2):524–539 **https://doi.org/10.1093/molbev/msz239**.

Liberti et al., 2021 Liberti A, Natarajan O, Atkinson CGF, Sordino P, Dishaw LJ. Reflections on the use of an invertebrate chordate model system for studies of gut microbial immune interactions. *Frontiers in Immunology*. 2021;12 **https://doi.org/10.3389/fimmu.2021.642687**.

Liu et al., 2018 Liu H, Stephens TG, González-Pech RA, et al. Symbiodinium genomes reveal adaptive evolution of functions related to coral-dinoflagellate symbiosis. *Communications biology*. 2018;1:95 **https://doi.org/10.1038/s42003-018-0098-3**.

Marulanda-Gomez et al., 2024 Marulanda-Gomez AM, Ribes M, Franzenburg S, Hentschel U, Pita L. Transcriptomic responses of Mediterranean sponges upon encounter with symbiont microbial consortia. *BMC Genomics*. 2024;25 **https://doi.org/10.1186/s12864-024-10548-z**.

Mazel et al., 2018 Mazel F, Davis KM, Loudon A, Kwong WK, Groussin M, Parfrey LW. Is host filtering the main driver of phylosymbiosis across the tree of life?. *mSystems*. 2018;3 **https://doi.org/10.1128/MSYSTEMS.00097-18**.

McNulty et al., 2017 McNulty SN, Tort JF, Rinaldi G, et al. Genomes of *Fasciola hepatica* from the Americas reveal colonization with Neorickettsia endobacteria related to the agents of Potomac horse and human sennetsu fevers. *PLoS Genetics*. 2017;13 **https://doi.org/10.1371/journal.pgen.1006537**.

Midha et al., 2022 Midha A, Jarquín-Díaz VH, Ebner F, et al. Guts within guts: The microbiome of the intestinal helminth parasite Ascaris suum is derived but distinct from its host. *Microbiome*. 2022;10 **https://doi.org/10.1186/s40168-022-01399-5**.

Midha et al., 2021 Midha A, Ebner F, Schlosser-Brandenburg J, Rausch S, Hartmann S. Trilateral relationship: Ascaris, microbiota, and host cells. *Trends in Parasitology*. 2021;37(3):251–262 **https://doi.org/10.1016/J.PT.2020.09.002**.

Moitinho-Silva et al., 2017 Moitinho-Silva L, Nielsen S, Amir A, et al.... The sponge microbiome project. *GigaScience*. 2017;6 **https://doi.org/10.1093/gigascience/gix077**.

Muslim et al., 2024 Muslim A, Aazmi S, Er YX, Shahrizal S, Lim YAL. Ascaris lumbricoides harbors a distinct gut microbiota profile from its human host: Preliminary insights. *Food and Waterborne Parasitology*. 2024;34 **https://doi.org/10.1016/j.fawpar.2024.e00223**.

Needham et al., 2019 Needham DM, Yoshizawa S, Hosaka T, et al.... A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators. *Proceedings of the National Academy of Sciences of the United States of America*. 2019;116(41):20574–20583 **https://doi.org/10.1073/pnas.1907517116**.

Needham et al., 2022 Needham D, Poirier C, Bachy C, George E, Wilken S, Yung C. The microbiome of a bacterivorous marine choanoflagellate contains a resource-demanding obligate bacterial associate. *Nat Microbiol*. 2022;7(9):1466–1479 **https://doi.org/10.1038/s41564-022-01174-0**.

Oulhen et al., 2016 Oulhen N, Schulz BJ, Carrier TJ. English translation of Heinrich Anton de Bary's 1878 speech, 'Die Erscheinung der Symbiose' ('De la symbiose'). *Symbiosis*. 2016;69(3):131–139 **https://doi.org/10.1007/s13199-016-0409-8**.

Palevich et al., 2023 Palevich N, Maclean PH, Carbone V, Jauregui R, Umair S. Multi-omic profiling, structural characterization, and

potent inhibitor screening of evasion-related proteins of a parasitic nematode, haemonchus contortus, surviving vaccine treatment. *Biomedicines*. 2023;11 **https://doi.org/10.3390/biomedicines11020411**.

Pittis and Gabaldón, 2016 Pittis AA, Gabaldón T. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature*. 2016;531(7592):101–104 **https://doi.org/10.1038/nature16941**.

Pushpakumara et al., 2023 Pushpakumara BLDU, Tandon K, Willis A, Verbruggen H. The bacterial microbiome of the coral skeleton algal symbiont ostreobium shows preferential associations and signatures of phylosymbiosis. *Microbial Ecology*. 2023;86(3):2032–2046 **https://doi.org/10.1007/s00248-023-02209-7**.

Pusterla et al., 2003 Pusterla N, Johnson EM, Chae JS, Madigan JE. Digenetic trematodes, *Acanthatrium* sp and *Lecithodendrium* sp., as vectors of *Neorickettsia risticii*, the agent of Potomac horse fever. *Journal of Helminthology*. 2003;77(4):335–339 **https://doi.org/10.1079/joh2003181**.

Qin et al., 2023 Qin M, Jiang L, Qiao G, Chen J. Phylosymbiosis: The eco-evolutionary pattern of insect–symbiont interactions. *International Journal of Molecular Sciences*. 2023;24(21):15836 **https://doi.org/10.3390/ijms242115836**.

Reyes-Prieto et al., 2007 Reyes-Prieto A, Weber APM, Bhattacharya D. The origin and establishment of the plastid in algae and plants. *Annual Review of Genetics*. 2007;41:147–168 **https://doi.org/10.1146/annurev.genet.41.110306.130134**.

Rinaldi et al., 2024 Rinaldi G, Paz Meseguer C, Cantacessi C, Cortés A. Form and function in the Digenea, with an emphasis on host–parasite and parasite–bacteria interactions. *Advances in*

Experimental Medicine and Biology. 2024;1454:3–45 **https://doi.org/10.1007/978-3-031-60121-7_1**.

Ritchie, 2022 Ritchie H. (2022). How many species are there? Published online at OurWorldinData.org. Retrieved from: **https://ourworldindata.org/how-many-species-are-there** [Online Resource].

Rivera et al., 1998 Rivera MC, Jain R, Moore JE, Lake JA. Genomic evidence for two functionally distinct gene classes. *Proceedings of the National Academy of Sciences of the United States of America*. 1998;95(11):6239–6244 **https://doi.org/10.1073/pnas.95.11.6239**.

Rivera and Lake, 1992 Rivera MC, Lake JA. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science (New York, N.Y.)*. 1992;257(5066):74–76 **https://doi.org/10.1126/science.1621096**.

Rivera and Lake, 2004 Rivera MC, Lake JA. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature*. 2004;431(7005):152–155 **https://doi.org/10.1038/nature02848**.

Rook, 2009 Rook GAW. Review series on helminths, immune modulation and the hygiene hypothesis: The broader implications of the hygiene hypothesis. *Immunology*. 2009;126(1):3–11 **https://doi.org/10.1111/j.1365-2567.2008.03007.x**.

Rosenberg and Zilber-Rosenberg, 2011 Rosenberg E, Zilber-Rosenberg I. Symbiosis and development: The hologenome concept. *Birth Defects Research Part C: Embryo Today: Reviews*. 2011;93(1):56–66 **https://doi.org/10.1002/bdrc.20196**.

Rothschild et al., 2018 Rothschild D, Weissbrod O, Barkan E, et al.... Environment dominates over host genetics in shaping human gut

microbiota. *Nature*. 2018;555(7695):210–215 **https://doi.org/10.1038/nature25973**.

Schopf, 1993 Schopf JW. Microfossils of the early Archean apex chert: New evidence of the antiquity of life. *Science (New York, N.Y.)*. 1993;260(5108):640–646 **https://doi.org/10.1126/science.260.5108.640**.

Seymour et al., 2017 Seymour JR, Amin SA, Raina JB, Stocker R. Zooming in on the phycosphere: The ecological interface for phytoplankton-bacteria relationships. *Nature Microbiology*. 2017;2 **https://doi.org/10.1038/nmicrobiol.2017.65**.

Sinnathamby et al., 2018 Sinnathamby G, Henderson G, Umair S, et al. The bacterial community associated with the sheep gastrointestinal nematode parasite Haemonchus contortus. *PLoS One*. 2018;13(2):e0192164 **https://doi.org/10.1371/journal.pone.0192164**.

Sironi et al., 1995 Sironi M, Bandi C, Sacchi L, Sacco BD, Damiani G, Genchi C. Molecular evidence for a close relative of the arthropod endosymbiont Wolbachia in a filarial worm. *Molecular and Biochemical Parasitology*. 1995;74(2):223–227 **https://doi.org/10.1016/0166-6851(95)02494-8**.

Slatko et al., 2010 Slatko BE, Taylor MJ, Foster JM. The Wolbachia endosymbiont as an anti-filarial nematode target. *Symbiosis*. 2010;51(1):55–65 **https://doi.org/10.1007/s13199-010-0067-1**.

Spang et al., 2015 Spang A, Saw JH, Jørgensen SL, et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*. 2015;521(7551):173–179 **https://doi.org/10.1038/nature14447**.

Speth et al., 2024 Speth DR, Zeller LM, Graf JS, Overholt WA, Küsel K, Milucka J. Genetic potential for aerobic respiration and denitrification in globally distributed respiratory endosymbionts.

Nature Communications. 2024;15 **https://doi.org/10.1038/s41467-024-54047-x**.

Stork, 2018 Stork NE. How many species of insects and other terrestrial arthropods are there on earth?. *Annual Review of Entomology*. 2018;63:31–45 **https://doi.org/10.1146/annurev-ento-020117-043348**.

Telford et al., 2015 Telford MJ, Budd GE, Philippe H. Phylogenomic insights into animal evolution. *Current Biology*. 2015;25(19):R876 **https://doi.org/10.1016/j.cub.2015.07.060**.

Thomas et al., 2016 Thomas T, Moitinho-Silva L, Lurgi M, et al.... Diversity, structure and convergent evolution of the global sponge microbiome. *Nature Communications*. 2016;7:11870 **https://doi.org/10.1038/ncomms11870**.

Tinker and Ottesen, 2020 Tinker KA, Ottesen EA. Phylosymbiosis across deeply diverging lineages of omnivorous cockroaches (order blattodea). *Applied and Environmental Microbiology*. 2020;86 **https://doi.org/10.1128/AEM.02513-19**.

van Opstal and Bordenstein, 2019 van Opstal EJ, Bordenstein SR. Phylosymbiosis impacts adaptive traits in nasonia wasps. *mBio*. 2019;10 **https://doi.org/10.1128/mBio.00887-19**.

Vaughan et al., 2012 Vaughan JA, Tkach VV, Greiman SE. Neorickettsial endosymbionts of the Digenea: Diversity, transmission and distribution. *Advances in Parasitology*. 2012;79:253–297 **https://doi.org/10.1016/B978-0-12-398457-9.00003-2**.

Vosseberg et al., 2024 Vosseberg J, van Hooff JJE, Köstlbacher S, Panagiotou K, Tamarit D, Ettema TJG. The emerging view on the origin and early evolution of eukaryotic cells. *Nature*. 2024;633(8029):295–305 **https://doi.org/10.1038/s41586-024-07677-6**.

Werren et al., 2008 Werren JH, Baldo L, Clark ME. Wolbachia: Master manipulators of invertebrate biology. *Nature Reviews Microbiology*. 2008;6(10):741–751 **https://doi.org/10.1038/nrmicro1969**.

Wilkinson et al., 2021 Wilkinson BH, Ivany LC, Drummond CN. Estimating vertebrate biodiversity using the tempo of taxonomy —A view from Hubbert's peak. *Biological Journal of the Linnean Society*. 2021;134(2):402–422 **https://doi.org/10.1093/biolinnean/blab080**.

Williams and Fontaine, 2024 Williams CE, Fontaine SS. Commentary: The microbial dependence continuum: Towards a comparative physiology approach to understand host reliance on microbes. *Comparative Biochemistry and Physiology -Part A: Molecular and Integrative Physiology*. 2024;296 **https://doi.org/10.1016/j.cbpa.2024.111690**.

Williams and Embley, 2014 Williams TA, Embley TM. Archaeal "dark matter" and the origin of eukaryotes. *Genome Biology and Evolution*. 2014;6(3):474–481 **https://doi.org/10.1093/gbe/evu031**.

Williams et al., 2013 Williams TA, Foster PG, Cox CJ, Embley TM. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature*. 2013;504(7479):231–236 **https://doi.org/10.1038/nature12779**.

World Health Organisation, 2022 World Health Organisation. (2022). Soil-transmitted Helminth Infections. **https://www.who.int/news-room/fact-sheets/detail/soil-transmitted-helminth-infections**.

Woznica and Richards, 2024 Woznica A, Richards TA. What choanoflagellates can teach us about symbiosis. *PLoS Biology*. 2024;22(4):e3002561 **https://doi.org/10.1371/journal.pbio.3002561**.

Wylezich et al., 2019 Wylezich C, Belka A, Hanke D, Beer M, Blome S, Höper D. Metagenomics for broad and improved parasite detection: A proof-of-concept study using swine faecal samples. *International Journal for Parasitology*. 2019;49(10):769–777 **https://doi.org/10.1016/j.ijpara.2019.04.007**.

Yoon et al., 2004 Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D. A molecular timeline for the origin of photosynthetic eukaryotes. *Molecular Biology and Evolution*. 2004;21(5):809–818 **https://doi.org/10.1093/molbev/msh075**.

Youngblut et al., 2019 Youngblut ND, Reischer GH, Walters W, et al. Host diet and evolutionary history explain different aspects of gut microbiome diversity among vertebrate clades. *Nature Communications*. 2019;10(1):2200 **https://doi.org/10.1038/s41467-019-10191-3**.

Zilber-Rosenberg and Rosenberg, 2008 Zilber-Rosenberg I, Rosenberg E. Role of microorganisms in the evolution of animals and plants: The hologenome theory of evolution. *FEMS Microbiology Reviews*. 2008;32(5):723–735 **https://doi.org/10.1111/j.1574-6976.2008.00123.x**.

CHAPTER 5

Hypothesis testing of metagenomic data

Rohan Mishra[1] and William D. Shannon[1,2], [1]BioRankings, St. Louis, MO, United States, [2]Washington University School of Medicine, St. Louis, MO, United States

Abstract

This chapter introduces the steps and considerations for hypothesis testing and experimental design for microbiome/metagenomic data. Experimental design is a key step in any clinical trial. Without proper experimental design an experiment can become over- or underpowered, with both cases leading to wasted money due to running too large an experiment or the need for a follow-up experiment. Choosing the right hypothesis test for your data is one of the most important factors when designing your experiment. These principles are presented with three scenarios as follows: (1) calculating diversity between groups, (2) finding differences between groups in one taxon, and (3) detecting differences between groups across all taxa. For the third scenario the Dirichlet-multinomial distribution is introduced, allowing for parametric statistical testing of the entire microbiome. Testing the entire microbiome allows for studies with greater statistical power and fewer Type I and Type II errors.

Keywords

Metagenomic data; Shannon diversity; compositional data; sample size calculation; power calculation; parametric hypothesis testing; experimental design; Dirichlet-multinomial distribution

Introduction

Hypothesis testing is the statistical procedure to decide if the data collected in an experiment provides enough evidence to reject a *null hypothesis*—that data from different groups comes from the same distribution (i.e., no difference in the data)—in favor of an *alternative hypothesis*—that the data comes from different distributions. To start with a simple example consider two groups of data whose measurements both follow a normal distribution, with differing means and the same variance. **Fig. 5.1** below shows a histogram of these two groups of data, with the possible values on the x-axis and the frequency of each value on the y-axis. The histogram shows that these two groups are visually different, with little overlap between them.

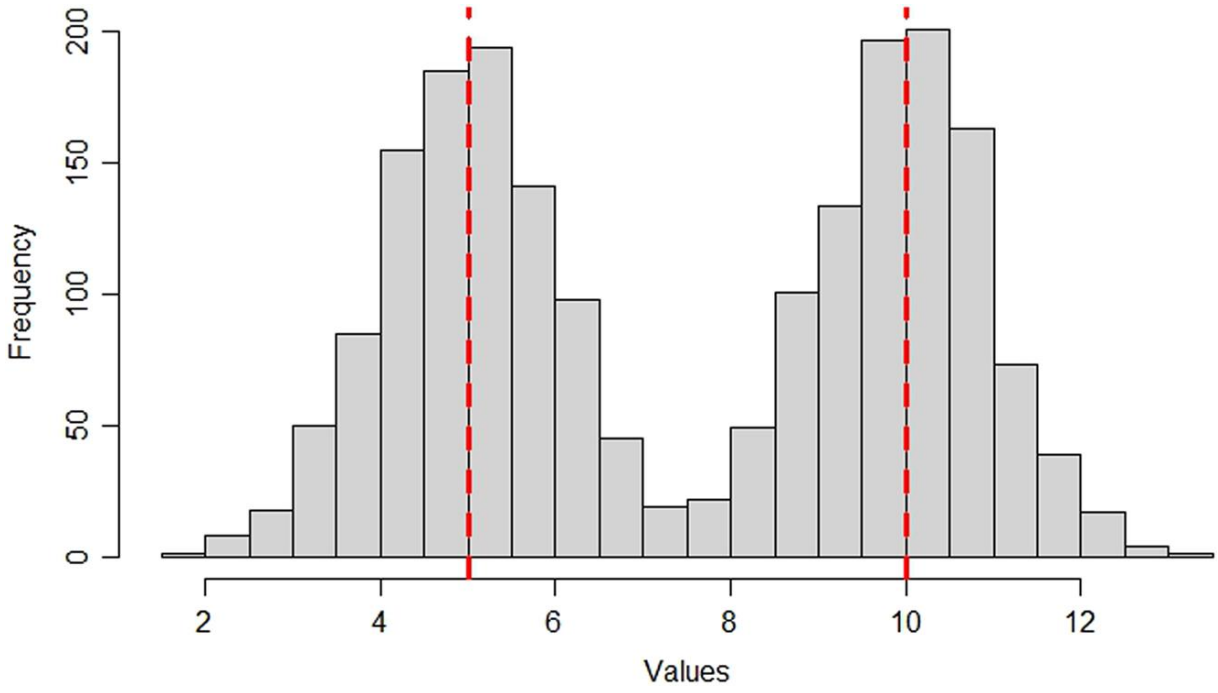


FIGURE 5.1 Hypothetical normal distributions of two groups with different means but the same variance. The red dashed lines represent the means of each group, and the distance between the dashed lines is the effect size between the two groups.

Follow for extended description

To decide if the distributions are the same or different in the two groups, a formal process is followed. Firstly formulate the problem and identify variables to be measured and compared—this step is already complete in the example above. Then identify a null and alternative hypothesis. From there four variables must be determined or calculated as follows:

- **Power:** a measure of the ability of an experimental design/hypothesis test to detect an effect when it is truly present, that is, a power of 0.8 (or 80%) means an experiment would detect a true difference 80% of the time
- **Significance level:** the threshold for determining statistical significance, that is, $P<.05$
- **Effect size:** the difference between two or more groups of data, that is, the distance between means
- **Sample size:** the number of subjects or samples in the study

After these four variables are calculated, an appropriate test statistic can be identified. Finally the value of the test statistic is calculated. This test statistic is associated with a *P*-value, which determines if the results of the study are statistically significant.

Fig. 5.2 shows the power for $P<.05$ significance levels for different sample sizes and three different effect sizes. For this example an effect size of 0.8 is considered large, 0.5 is considered medium, and 0.2 is considered small. The classic reference for sample size and power calculations is Muller & Cohen (1989), and a more recent introductory article is Jones et al. (2003). If the effect size is large, the distance between the groups is large, and 26 subjects randomized to both groups are needed to achieve 80% power (where the power curve intersects the horizontal line at 80% power). With a medium effect size 64 subjects per group are needed, and for a small effect size, 400 subjects are needed (not shown on plot).

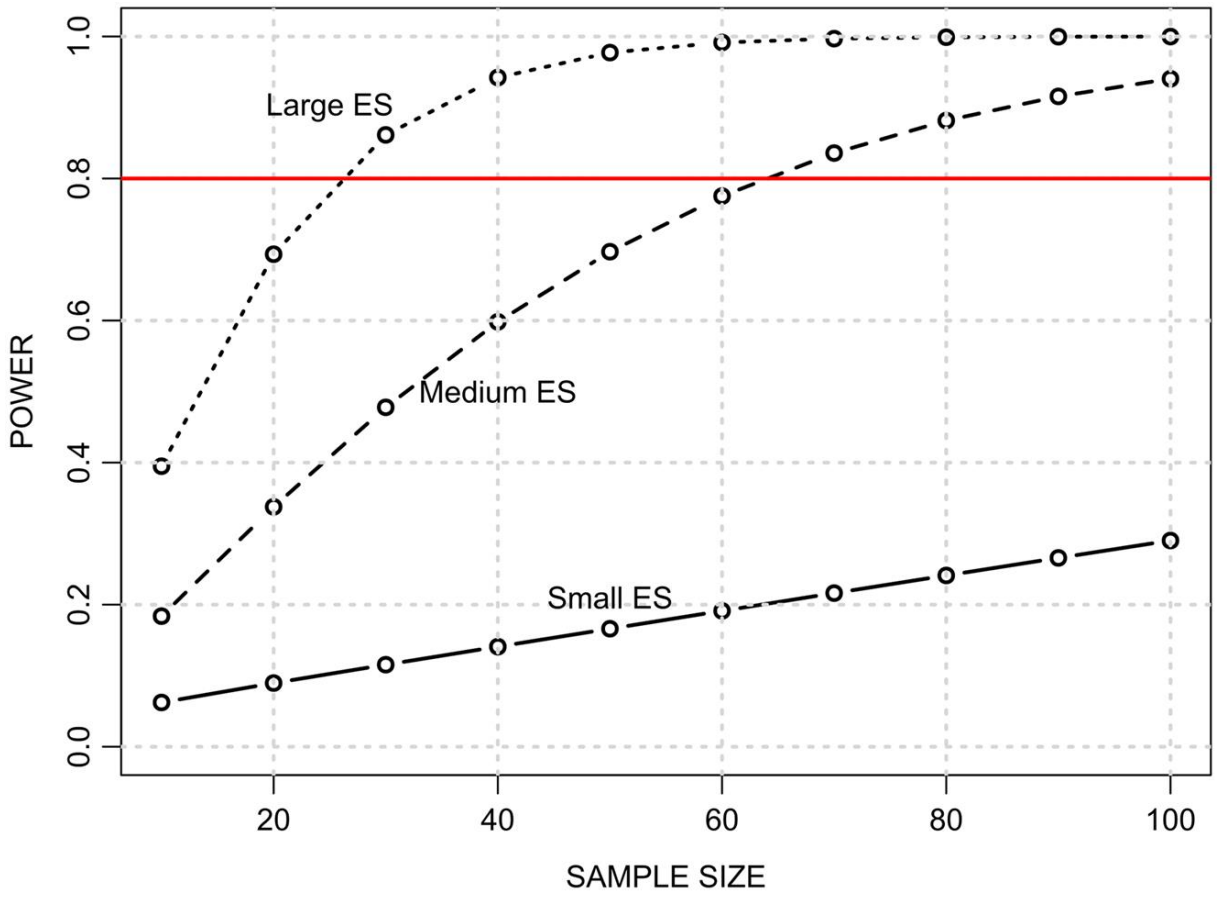


FIGURE 5.2 Power curve for a two-sided *t*-test with SD = 1 and $P \leq .05$ significance level for small, medium, and large effect sizes. The red line is for power = 0.8 (or 80%). The trend across the three lines shows a larger effect size (larger distance between distributions), and fewer samples are needed for the same power.

Follow for extended description

The approach presented previoulsy applies to all statistical hypothesis testing and design of experiments, including microbiome studies. In this chapter methods are presented for comparing microbiome diversity across groups, testing the proportion of counts for a single taxon, and testing the entire composition of microbiome samples across groups. For the scope of this chapter a taxon is a unique taxonomic name for an organism at any rank, such as class, family, and species. For ease of Page 97 presentation these

examples are limited to two group comparisons, but the ideas extend to more than two groups, though the statistical test may differ.

Metagenomic data

Microbiome data is formatted as shown in **Table 5.1**. The entries in the table are the number of sequence reads or counts from a sample's next generation sequencing (NGS) assay assigned to a particular taxon designated X_{ik} for subject "i" and taxon "k". The total number of reads for a sample is calculated by adding the counts across the row and is denoted by the symbol X_{i*}, where the * notation indicates summing. The total reads per sample are shown in the far-right column. Similarly the total number of reads per taxon is calculated by adding the counts down the columns and is denoted by the symbol X_{*k}. The total reads per taxon are shown in the bottom row. The value X_{**} in the bottom right table cell is the total number of sequence reads from the entire experiment summing across all taxa and subjects. Page 98

Table 5.1

Sample	Taxon 1	Taxon 2	...	Taxon k	Total reads per sample
1	$X_{1,1}$	$X_{1,2}$			$X_{1,*}$
2	$X_{2,1}$	$X_{2,2}$			
...					
i				$X_{i,k}$	$X_{i,*}$
Total reads per Taxon	$X_{*,1}$			$X_{*,K}$	$X_{*,*}$

The total number of reads per sample is a random variable and can vary widely across samples. This variability can result in apparent differences across groups that are not real. To avoid this many analysts will adjust the read counts by standardizing to make them equal across samples. One approach is to turn each count X_{ik} into a proportion by dividing it by the samples total read count X_{i*}. Since investigators choose to transform their data differently, (Legendre & Legendre, 2012), we will not provide an exhaustive list of formula here but do encourage the reader to carefully think about how the transformations can affect their data and the loss of information from

these changes (McMurdie & Holmes, 2014). This loss can greatly increase the false negative rates and by converting counts with millions of total reads across all taxa to 100, since percentages are a number out of 100. Due to this transformation, the sample size within each sample is reduced, reducing the power of the experiment. Ultimately the analyst will miss important taxa differences that are concealed by confidence intervals that are too wide due to the underpowered analysis.

Compare diversity across groups

Shannon diversity index, a measure of diversity of species/taxa in a given community, is calculated for each sample from a stool microbiome study of 136 Vervet monkeys comparing the impact of a typical American diet (TAD) versus a commercially available monkey feed (CHOW) (unpublished, personal communication). **Fig. 5.3** shows the distribution of Shannon diversity indexes for each group. A *t*-test rejected the *null hypothesis* that the TAD and CHOW diets have the same distribution for Shannon diversity (P = 2.646e-10). Note that the Page 99 distributions are slightly skewed, so a Wilcoxon test was conducted giving P = 2.869e-09, leading to the same conclusion.

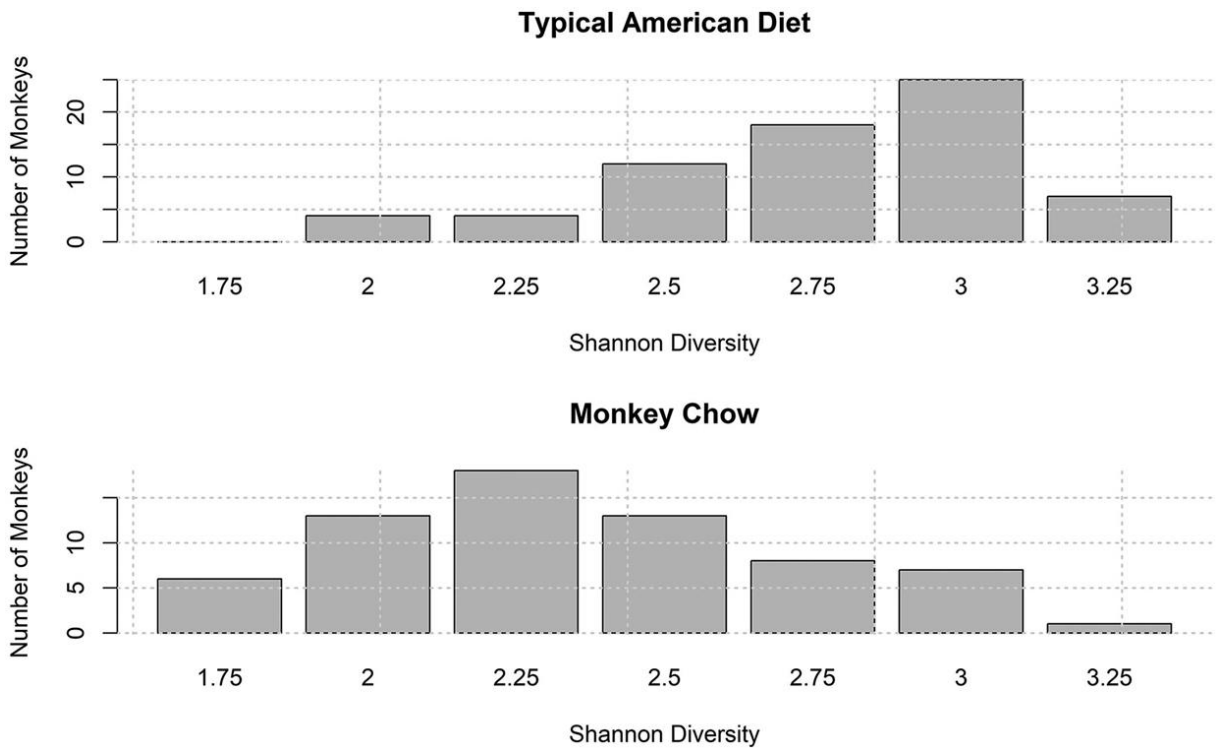


FIGURE 5.3 Shannon diversity index distributions for the gut microbiome of Vervet monkeys fed TAD and CHOW diet. The x-axis is the possible Shannon diversity values. The y-axis is the frequency of each diversity value within the group. A higher bar means that this Shannon diversity value is more frequent.

Follow for extended description

The Shannon diversity values for the TAD-fed monkeys had a mean = 2.78 and standard deviation = 0.324, and the CHOW-fed monkeys had a mean = 2.36 and standard deviation = 0.378. From this data the effect size is the difference between the means (effect size = 2.78–2.36), and calculating the power or sample size can be done in almost any statistical software package. For example the function in R **power.t.test** generates the power curves, as shown in **Fig. 5.4**, showing that for an experiment with 80% power, eight monkeys are needed for a significance level of 0.05, and 12 monkeys are needed for a significance level of 0.01.

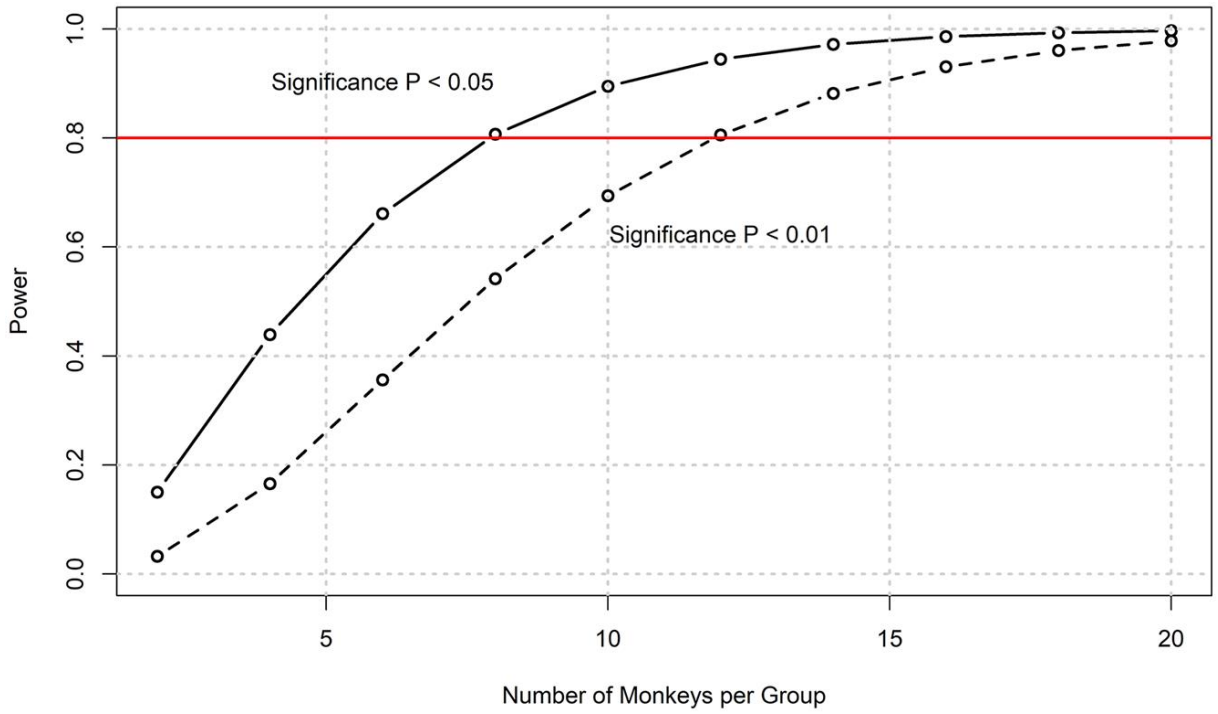


FIGURE 5.4 Power curves for testing mean Shannon diversities to confirm the results in **Fig. 5.3**. Two curves are shown for $P<.05$ and .01 indicating that $N = 8$ or 12 monkeys per group, respectively, are enough to obtain 80% power.

Follow for extended description

Comparing a single taxon across groups

Consider the problem of testing a single taxon across groups. It is important to specify the taxon *a priori* to ensure that the researcher is not looking at the data to see which taxa seem most different and then running tests to prove that what they saw is real. Testing data that has been looked at and selected based on apparent differences is called p-hacking, data dredging, or cherry-picking, leading to an inflation of false positives (Dwan et al., 2013). Page 100

Analysts transform the read counts of taxon k for sample i into a proportion p_{ik} by dividing its read counts $X_{i,k}$ by the sample's total read counts $X_{i,*}$ (i.e., $p_{ik} = X_{i,k}/X_{i,*}$). The proportions are then compared across groups using a *t*-test which assumes normality. This approach is incorrect, since proportions do not

meet the assumption of normality, rendering the *t*-test invalid. This is especially the case in microbiome data, due to the inherent sparseness of the data. Sparseness refers to the tendency of microbiome datasets to contain more zero values than nonzero values. Additionally proportions have unequal variances across the data, especially when sparseness is considered. The assumption of constant variance, called homoscedasticity, is required in almost all statistical analyses, regardless of domain. To correctly analyze taxon proportions, the analyst may consider the arcsine transformation:

$$y = \sin^{-1}(\sqrt{x})$$

▶ MathML Presentaion

to induce homoscedasticity (equal variance) and possibly make the data appear normally distributed (Steel et al., 1996). With this transformation the *t*-test is the correct test.

We suggest two alternatives to using *t*-tests to compare proportions across groups for a single taxon: The first option is to use analysis of covariance (ANCOVA), where the raw read count for the taxon of Page 101 interest is regressed onto both the total read count for the sample and the group membership. This adjusts for differences in total read counts across samples and is a more powerful method than transforming the data into proportions and using a *t*-test. This approach is not discussed further here, but the reader is encouraged to explore this option. The second option is to make a present versus absent comparison in samples. If a taxon has one or more reads, it is present in the sample, otherwise it is absent. Other thresholds can be used such as the taxon has five or more reads. This is useful to compare rare taxa. To illustrate we compared the rate of *Streptococcus* in samples from the left retroauricular crease and the gastrointestinal tract from the Human Microbiome Project (HMP) data (Human Microbiome Project Consortium 2012). **Table 5.2** shows the presence/absence counts and percentage within a

body site. To test the null hypothesis that there are no differences in the number of samples with this taxon between the two sites, a chi-square test is chosen, which returns a $P = 1.57 * 10^{-7}$. This provides enough evidence to reject the null hypothesis in favor of the alternative hypothesis: More samples have this taxon in the left crease (92%) compared with samples from the stool (65%).

Table 5.2

Contingency table showing counts of samples with (Present) and without (Absent) one or more sequence reads of *Streptococcus* sp.

Body Site	Present	Absent	Total
Left crease	167 (92%)	14 (8%)	181
Stool	135 (65%)	74 (35%)	209

To confirm this result in a follow-up study, 91 subjects per body site are needed to have 80% power for the observed effect size of 0.324 for the data in **Table 5.2** at a significance level of 0.05. This goes up to 131 subjects per body site for a significance level of 0.01. The power curves are shown in **Fig. 5.5**. An effect size of 0.324 for the chi-square test is generally assumed to be a medium effect size (Muller & Cohen, 1989).

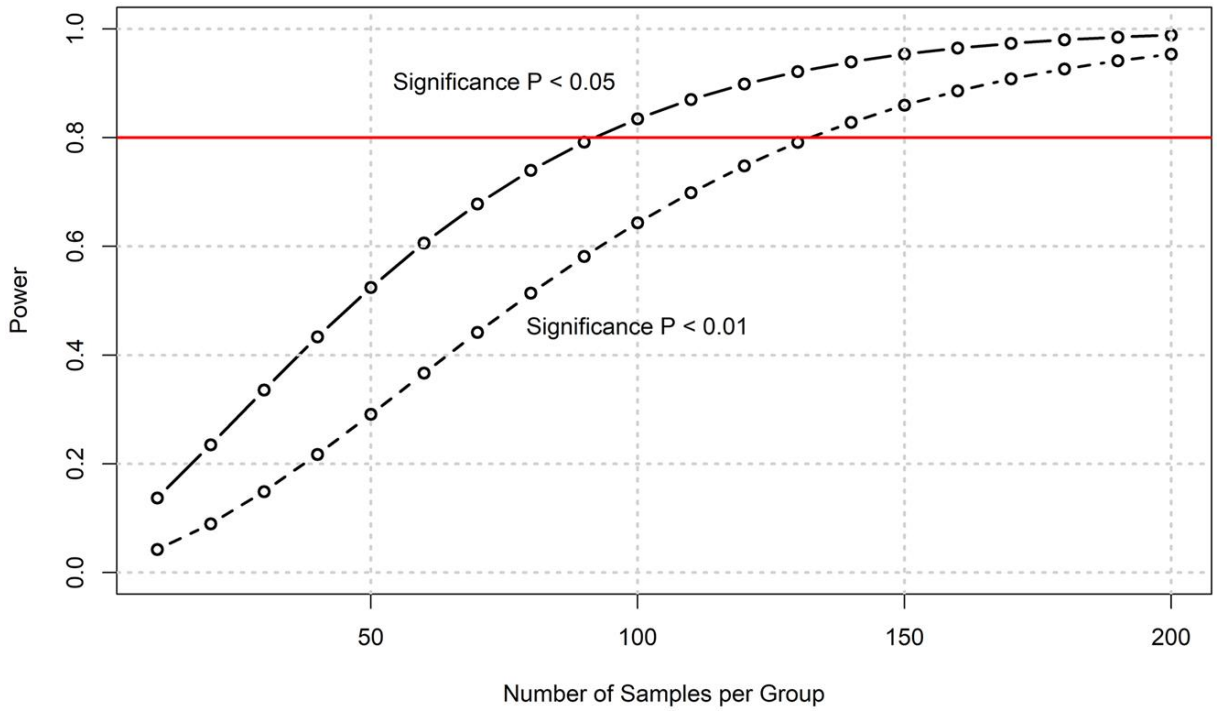


FIGURE 5.5 Power curves for present/absent contingency table comparing *Streptococcus* across two body sites. Two curves are shown for $P<.05$ and .01 indicating $N = 91$ or 131 per group, respectively, for 80% power.

Follow for extended description

A well-known problem of testing many taxa across groups separately is multiple testing. Consider the case where two taxa are separately tested with $P \leq .05$ as the level of significance for each test. In this case taxon A might be significant, taxon B might be significant, and both taxa A and B might be significant. The true Type I error (i.e., saying the taxon is different when it is not) or significance level must be adjusted by using one of several adjustments such as $1 - (1 - 0.05)^2 = 0.0975$. In other Page 102 words, while the analysts might be intended to test the hypothesis at $P<.05$, it is in fact testing them at $P<.0975$ increasing the probability of a false positive. For 10 separate tests the significance level is 0.4, meaning that for every taxon declared significant at $P = .05$, there is a 40% chance it could be a false positive. The impact of this on future research is wasted money on

preventable follow-up research to re-analyze the data. The literature on multiple testing is extensive, and the microbiome researcher should become familiar with it (Chen et al., 2017; Dunkler et al., 2011; Westfall & Young, 1993).

Analyzing one taxon at a time is easy to do but is known, as discussed above, to produce a lot of false positives. This approach also misses combinations of taxa that can drive clinical outcomes. In genetic epidemiology the combination of two or more markers is called epistasis and requires new tools for multivariate statistics (Culverhouse et al., 2004). In the next an approach to analyzing all taxa simultaneously is presented.

Compare all taxa across groups simultaneously

To analyze all the taxa simultaneously in a single test, the compositional structure of the data needs to be part of the analysis. For a microbiome study using NGS the number of reads that map to each taxon is assumed to be proportional to the relative abundance of the Page 103 taxon within the sample. For example, if a sample has 100 reads, each of those 100 reads is mapped to a taxon found in the sample. If 99 of the reads are Clostridia, then there can only be one more read for any other taxa. This is compositional data (Aitchison, 2003) and is a characteristic of microbiome data (Gloor et al., 2017). This means that changes in one taxon result in an opposite change across the other taxa.

The Dirichlet-multinomial (DM) model works with compositional data, specifically raw counts, as shown in **Table 5.1**. The DM model provides a framework for designing microbiome studies, including statistical testing to compare microbiome compositions across groups, effect size measurement to quantify how far apart groups are from each other, and power and sample size calculations (La Rosa et al., 2012). No *information-losing* transformations of the data such as standardization are needed (McMurdie & Holmes, 2014). The HMP open-source R software package contains the functions needed to

analyze microbiome compositional data and works on experiments with two or more groups (La Rosa et al., 2012).

Fig. 5.6 shows the taxa proportions and 95% confidence intervals for nine microbial genera and one group combining the rarer genera (called Page 104 "other") from the HMP throat and saliva samples. **Fig. 5.6** shows that *Streptococcus* sp. makes up almost 30% of the taxa in throat samples and about 18% in saliva samples. The other taxa show less variability between the two body sites. Each body site's raw taxon counts are modeled by the DM distribution with the relative abundances indicating the *mean* of the distribution, and the confidence intervals are calculated from the *variance* or *overdispersion* parameter of the distribution. The further apart these relative abundance curves are from each other, the larger the effect size. The DM hypothesis test for count data indicated the relative abundances of these genera over the two body sites (saliva and throat) are statistically significant ($P<.00001$). The relative abundances and confidence intervals of the two body sites in **Fig. 5.6** are far apart with a calculated effect size = 0.209 using the modified Cramer's Phi statistic (La Rosa et al., 2012). A power calculation using these abundance curve data, the effect size, and a significance level $P<.05$ requires $N = 34$ samples per body site to have 80% power, informing us of whether more measurements are needed for further validation of the test. **Table 5.3** shows the power analysis for this data using 1% ($P<.01$) and 5% ($P<.05$) significance levels. The table shows that to obtain an approximate power of 90%, at a 0.05 significance level, about 40 samples per body site will be needed, and more than 45 samples per body site will be needed at a 0.01 significance level. Also note that as the significance level gets smaller (i.e., $P \leq .01$ to reject the null hypothesis), more samples are needed to obtain the same power. Page 105

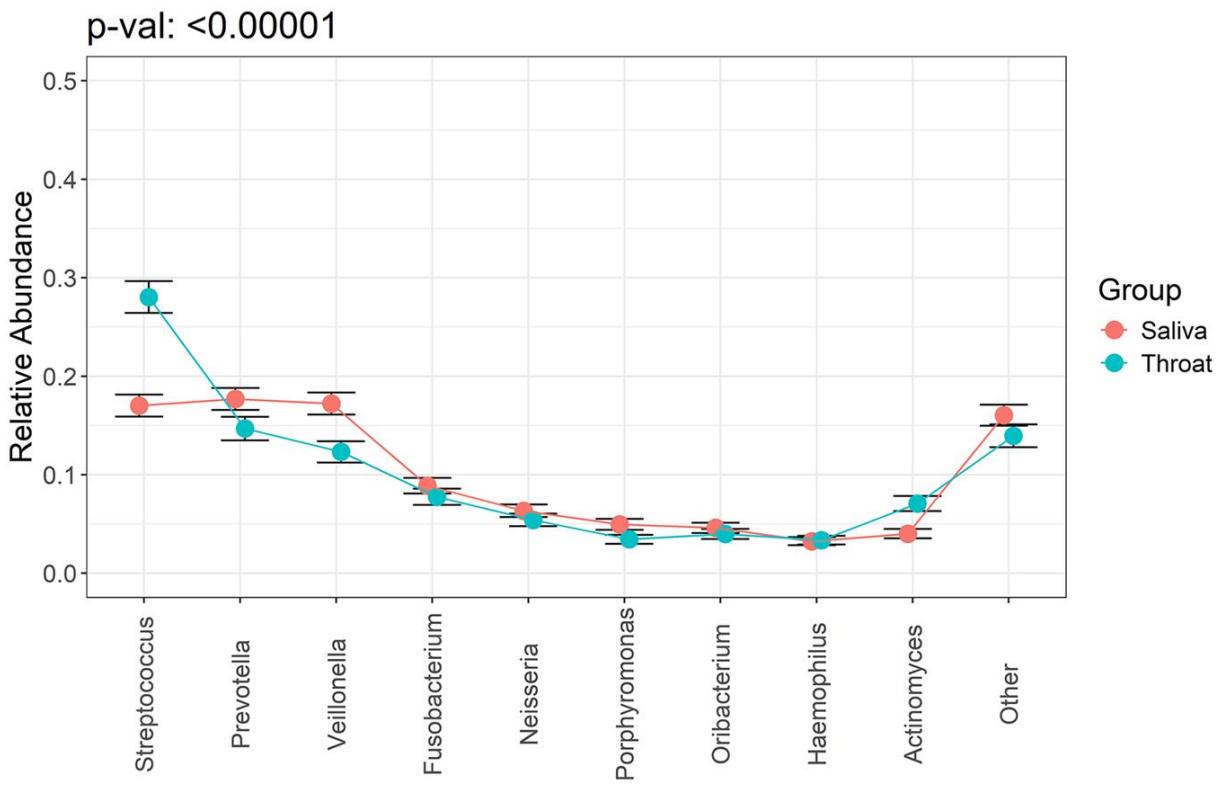


FIGURE 5.6 Plot of taxa relative abundance and 95% confidence intervals calculated using the DM model on raw taxa count data for saliva and throat. Saliva contained 194 samples, while throat contained 203 samples. The lines connecting abundances are included to make it easier to detect differences across groups.

Follow for extended description

Table 5.3

Power for comparing saliva and throat data at $P<.05$ and $P<.01$.

Number of samples	$P<.05$	$P<.01$
10	7%	2%
15	20%	9%
20	36%	20%
25	56%	35%
30	72%	53%
35	84%	68%
40	92%	80%
45	95%	89%
50	98%	94%

Parametric statistical models such as the DM provide powerful frameworks for data analysis based on more than a century of work in mathematical statistics and probability. Once a probability model is defined, the measurements are parameterized, which leads to formalized hypothesis testing and automatic *P*-values. For example the normal distribution has two parameters as follows: the mean and standard deviation. Once we have two groups of measurements that follow a normal distribution with a given mean and standard deviation, hypothesis testing can be done precisely because both groups fall under the same distribution. Similarly, by modeling microbiome data with the DM distribution, analysts can similarly perform hypothesis testing designed for the DM distribution. This also lets the analyst to perform power and sample size calculations and provide clear approaches to designing experiments that are more powerful and objective than any *ad hoc* algorithm. For example Nearing et al. (2022) showed that for 14 differential abundance

algorithms applied to 38 datasets, the results were vastly different which should be of concern to the microbiome data analyst.

Microbiome count data might be analyzed using a multinomial distribution. However this underestimates the variability in the data and so will result in lower power to detect significant differences. The reason the DM distribution is the right parametric model for microbiome data is that it measures the variability (technically called overdispersion) in the data correctly. Simulations showed power decreased from>99% for the DM to 72% for a multinomial model indicating the importance of measuring the overdispersion accurately (La Rosa et al., 2012). The R HMP package has a goodness-of-fit function to test if the DM distribution fits a microbiome read count dataset formatted, as shown in **Table 5.1**.

Discussion and future research

This chapter has introduced the application of hypothesis testing and power and sample size calculations to microbiome research. The goal in writing this chapter is to encourage microbiome investigators to consider these statistical topics early in their design of experiments to increase the probability of reaching the right conclusions about their data. The paper focused on topics related to: the format of microbiome taxonomic sequence counts data, the distribution of the data, the effect size indicating how far apart groups of microbiome samples are from Page 106 each other, and the formal parametric statistical tests and power/sample size calculations. Examples for comparing diversity measures across groups using the *t*-test, the statistical significance across groups under the DM distribution, and the rates of presence/absence of a taxon using the chi-square test were presented. An alternative to calculating and comparing taxon proportions across groups, ANCOVA, was suggested too. Short discussions were included about multiple testing adjustments (when each taxon is compared across groups separately) and the dangers in designing and analyzing data using the wrong data distribution.

Many investigators choose to use multiple testing (i.e., compare one taxon at a time across groups) and adjust the *P*-values using one of several different multiple-testing adjustment strategies (Westfall & Young, 1993). This approach to data analysis has the advantage of allowing the investigator to examine individual taxa to learn how they may impact a phenotype. The shortcoming of this approach is that it ignores the interactions or correlations among the taxa and treats them as independent, which is not true for metagenomic data. When the investigator is interested in discovery, this approach may be sufficient for generating hypotheses for further research. However, as the microbiome moves out of the basic science discovery phase to clinical translational research, with eventual development of biomarkers for disease and drug experiments under Food and Drug Administration guidelines, more formal statistics such as the multivariate DM test will almost surely be required.

In this chapter, because of space limitations, we have exclusively focused on parametric statistical testing and power calculations. However other methods such as PERMANOVA and ANOSIM permutation testing are frequently used (Xia & Sun, 2017). For investigators preferring to use the nonparametric approaches there is ample literature to guide them on their use.

As biostatisticians, developing and promoting the use of parametric models for analyzing biomedical data are almost always preferred since these generally have better properties than nonparametric and algorithmic models. For example parametric models are almost uniformly more efficient than nonparametric models, which mean statistically that in cases where 100 samples are required for a nonparametric model, the parametric model will have the same power with fewer samples leading to cost savings in experiments. There are two open problems to solve with the DM test. First, if the null hypothesis is Page 107 rejected for the alternative—that the groups are different—the question of which taxa are different arises. By rejecting the null hypothesis in a multivariate test the only conclusion drawn is that at least one of the taxa, if not all, is different. A *post hoc* test, analogous to those

used in analysis of variance, will tell us which taxa are different. Second, many microbiome studies are longitudinal, where samples are collected from the same subject at multiple times. This results in correlation of samples within subjects that must be considered to obtain accurate *P*-values and power/sample size estimates. A generalized DM distribution model can be used to solve this problem (Wilson & Chen, 2022).

References

Aitchison, 2003 Aitchison J. *The statistical analysis of compositional data* Caldwell, NJ: Blackburn Press; 2003;.

Chen et al., 2017 Chen SY, Feng Z, Yi X. A general introduction to adjustment for multiple comparisons. *Journal of Thoracic Disease*. 2017;9(6):1725–1729 **https://doi.org/10.21037/jtd.2017.05.34**.

Culverhouse et al., 2004 Culverhouse R, Klein T, Shannon W. Detecting epistatic interactions contributing to quantitative traits. *Genetic Epidemiology*. 2004;27(2):141–152 **https://doi.org/10.1002/gepi.20006**.

Dunkler et al., 2011 Dunkler D, Sánchez-Cabo F, Heinze G. Statistical analysis principles for omics data. *Methods in Molecular Biology*. 2011;719:113–131 **https://doi.org/10.1007/978-1-61779-027-0_5**.

Group Reporting Bias, 2013 Dwan K, Gamble C, Williamson PR, Kirkham JJ Group Reporting Bias. *Systematic review of the empirical evidence of study publication bias and outcome reporting bias – An updated review*. *PLoS One*. 2013;8 **https://doi.org/10.1371/journal.pone.0066844** United Kingdom.

Gloor et al., 2017 Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: And this is

not optional. *Frontiers in Microbiology*. 2017;8 **https://doi.org/10.3389/fmicb.2017.02224**.

Human Microbiome Project Consortium, 2012 Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*. 2012;486:207–214 **https://doi.org/10.1038/nature11234**.

Jones et al., 2003 Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. *Emergency Medicine Journal*. 2003;20(5):453–458 **https://doi.org/10.1136/emj.20.5.453**.

La Rosa et al., 2012 La Rosa PS, Brooks JP, Deych E, et al. Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS One*. 2012;7 **https://doi.org/10.1371/journal.pone.0052078**.

Legendre and Legendre, 2012 Legendre P, Legendre L. *Numerical Ecology* Amsterdam: Elsevier; 2012;.

McMurdie and Holmes, 2014 McMurdie PJ, Holmes S. Waste not, want not: Why rarefying microbiome data is inadmissible. *PLoS Computational Biology*. 2014;10 **https://doi.org/10.1371/journal.pcbi.1003531**.

Muller and Cohen, 1989 Muller K, Cohen J. Statistical power analysis for the behavioral sciences. *Technometrics*. 1989;31(4):499 **https://doi.org/10.2307/1270020**.

Nearing et al., 2022 Nearing JT, Douglas GM, Hayes MG, et al. Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications*. 2022;13 **https://doi.org/10.1038/s41467-022-28034-z**.

Steel et al., 1996 Steel R, Torrie J, Dickey D. *Principles and Procedures of Statistics: A Biometrical Approach* McGraw-Hill 1996;.

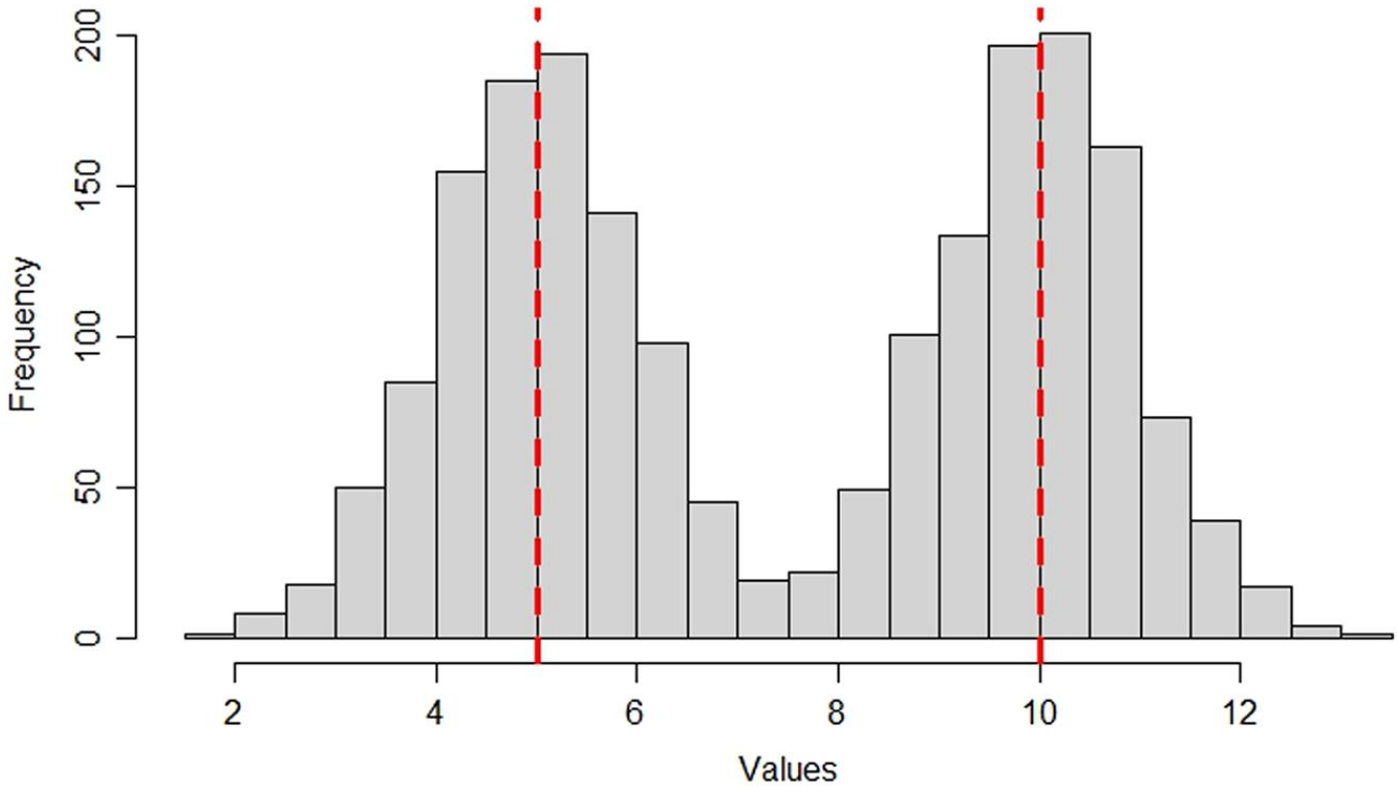
Westfall and Young, 1993 Westfall PH, Young SS. *Resampling-based multiple testing: Examples and methods for P-value adjustment*

New York: Wiley; 1993;.

Wilson and Chen, 2022 Wilson JR, Chen GSC. Dirichlet-multinomial model with varying response rates over time. *Journal of Data Science*. 2022;5:413–423 **https://doi.org/10.6339/JDS.2007.05(3).334**.

Xia and Sun, 2017 Xia Y, Sun J. Hypothesis testing and statistical analysis of microbiome. *Genes and Diseases*. 2017;4(3):138–148 **https://doi.org/10.1016/j.gendis.2017.06.001**.

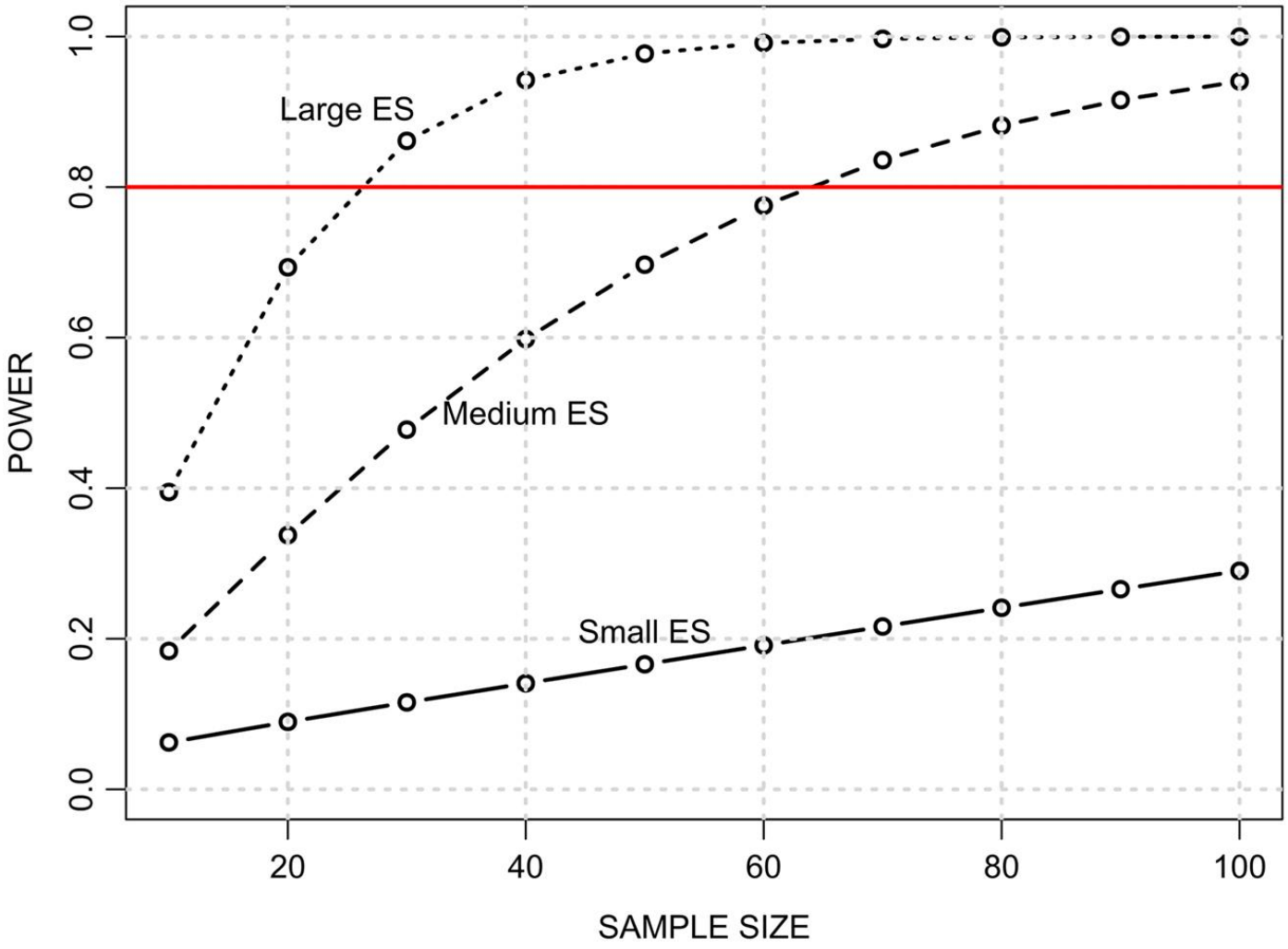
Description of Figure 5.1



The x-axis, labeled 'Values', extends from 2 to 12 in increments of 2. The y-axis, labeled 'Frequency', ranges from 0 to 200 in increments of 50. The distribution is bimodal, showing two distinct peaks. The first peak occurs around the value of 5, with a frequency of approximately 190. The second peak is centered around 10, also reaching a frequency of about 200. There are two vertical dashed lines on the graph. The first is positioned at approximately 5 on the x-axis and the second is at about 10, corresponding to the two peaks of the distribution. The detailed data points are as follows: For values equals 2, the frequency is 3. For values equals 3, the frequency is 10. For values equals 4, the frequency is 85. For values equals 5, the frequency is 190. For values equals 6, the frequency is 140. For values equals 7, the frequency is 95. For values equals 8, the frequency is 50. For values equals 9, the frequency is 135. For values equals 10, the frequency is 200. For values equals 11, the frequency is 165. For values equals 12, the frequency is 40.

Navigate back to Figure 5.1

Description of Figure 5.2

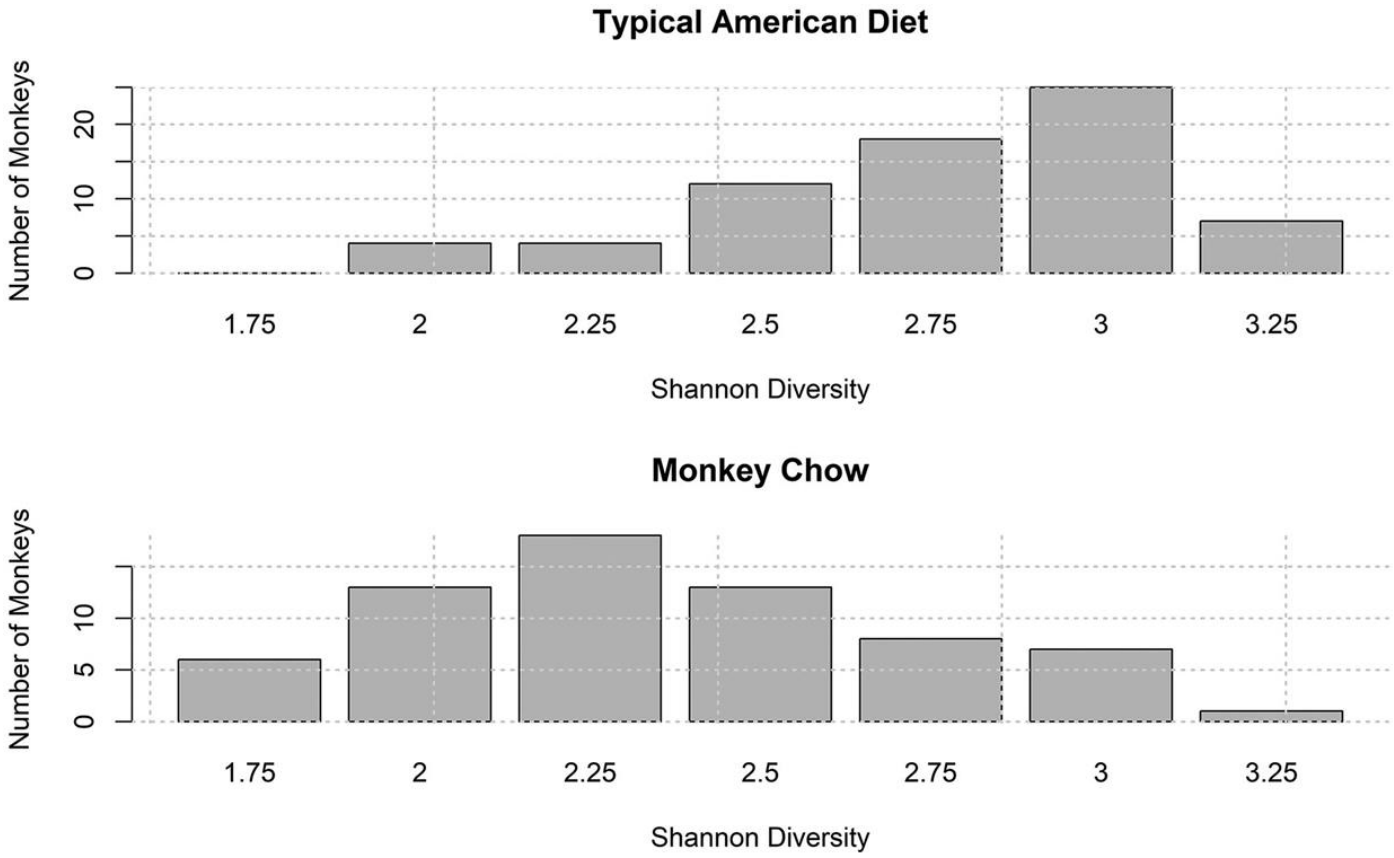


The graph displays the relationship between power and sample size for three different effect sizes (E S): Large, Medium and Small. The x-axis represents the sample size, ranging from 0 to 100 in increments of 20. The y-axis represents power, ranging from 0 to 1.0 in increments of 0.2. Three lines are plotted on the graph: 1. Large E S (dotted line): Starts at approximately (0, 0.4) and rapidly increases to nearly 1.0 power at a sample size of 60. It remains close to 1.0 for larger sample sizes. 2. Medium E S (dashed line): Begins around (0, 0.2) and increases more gradually, reaching 0.8 power at a sample size of about 60. It reaches 0.9 power at a sample size of 100. 3. Small E S (solid line): Starts near (0, 0.05) and increases very slowly, reaching only

about 0.3 power at a sample size of 100. A horizontal red line is drawn at the 0.8 power level, which intersects with the Medium ES line at approximately a sample size of 60.

Navigate back to Figure 5.2

Description of Figure 5.3

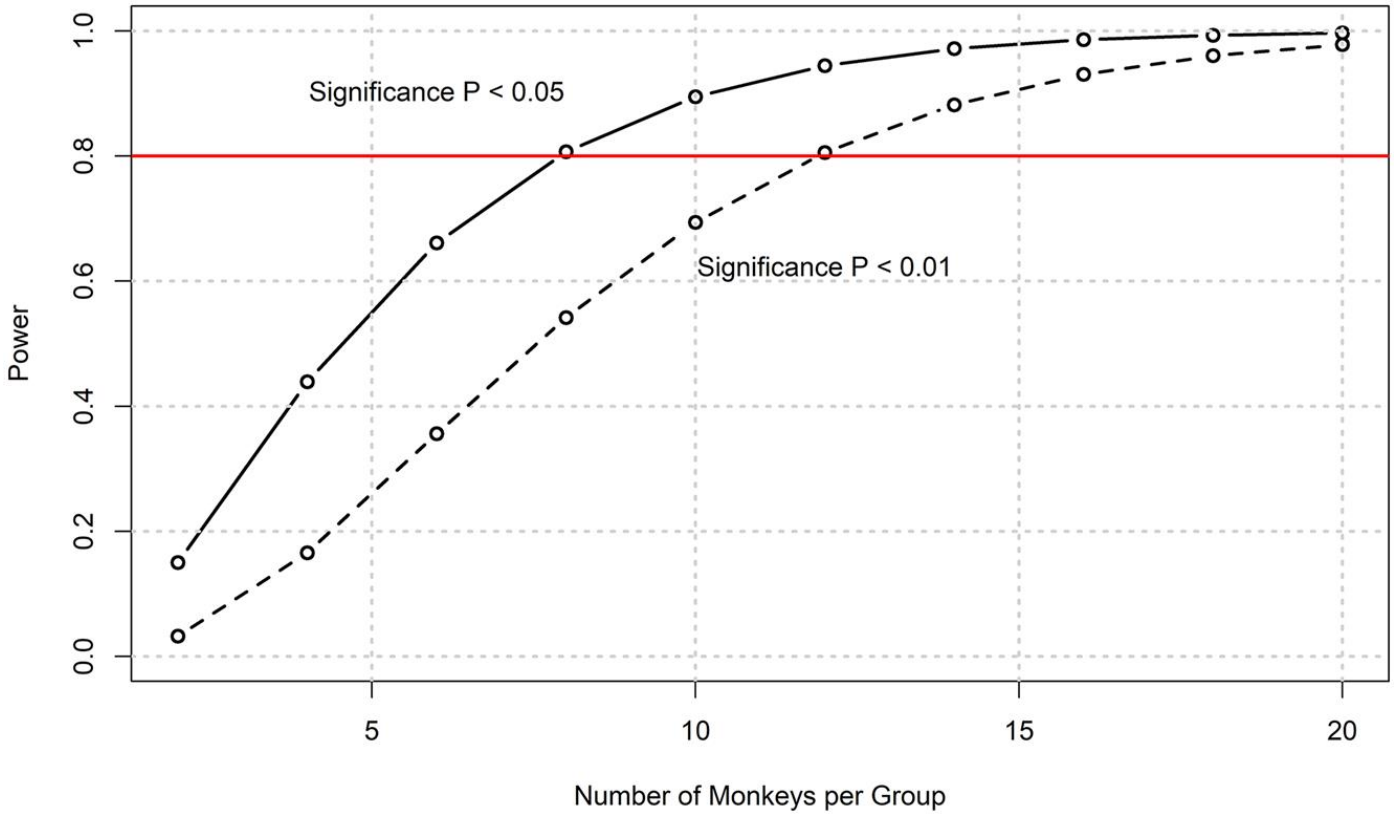


The image contains two histograms stacked vertically, both with the title 'Shannon Diversity' on the x-axis and 'Number of Monkeys' on the y-axis. The top histogram is labeled 'Typical American Diet'. The x-axis ranges from 1.75 to 3.25 in increments of 0.25. The y-axis ranges from 0 to 25 with markers at 0, 10 and 25. The distribution shows: 1.75 to 2.00: 0 monkeys 2.00 to 2.25: 3 monkeys 2.25 to 2.50: 5 monkeys 2.50 to 2.75: 12 monkeys 2.75 to 3.00: 20 monkeys 3.00 to 3.25: 25 monkeys 3.25 to 3.50: 8 monkeys. The bottom histogram is labeled 'Monkey Chow'. It has the same x-axis range of 1.75 to 3.25 in increments of 0.25. The y-axis ranges from 0 to 10 in increments of 5. The distribution shows: 1.75 to 2.00: 3 monkeys 2.00 to 2.25: 12 monkeys 2.25 to 2.50: 18 monkeys 2.50 to 2.75: 12 monkeys 2.75 to 3.00: 5 monkeys 3.00 to 3.25: 3 monkeys 3.25 to 3.50: 1 monkey The Typical American Diet histogram shows a right-skewed distribution with the peak at 3.00 to 3.25.

The Monkey Chow histogram shows a more symmetrical distribution with the peak at 2.25 to 2.50.

Navigate back to Figure 5.3

Description of Figure 5.4

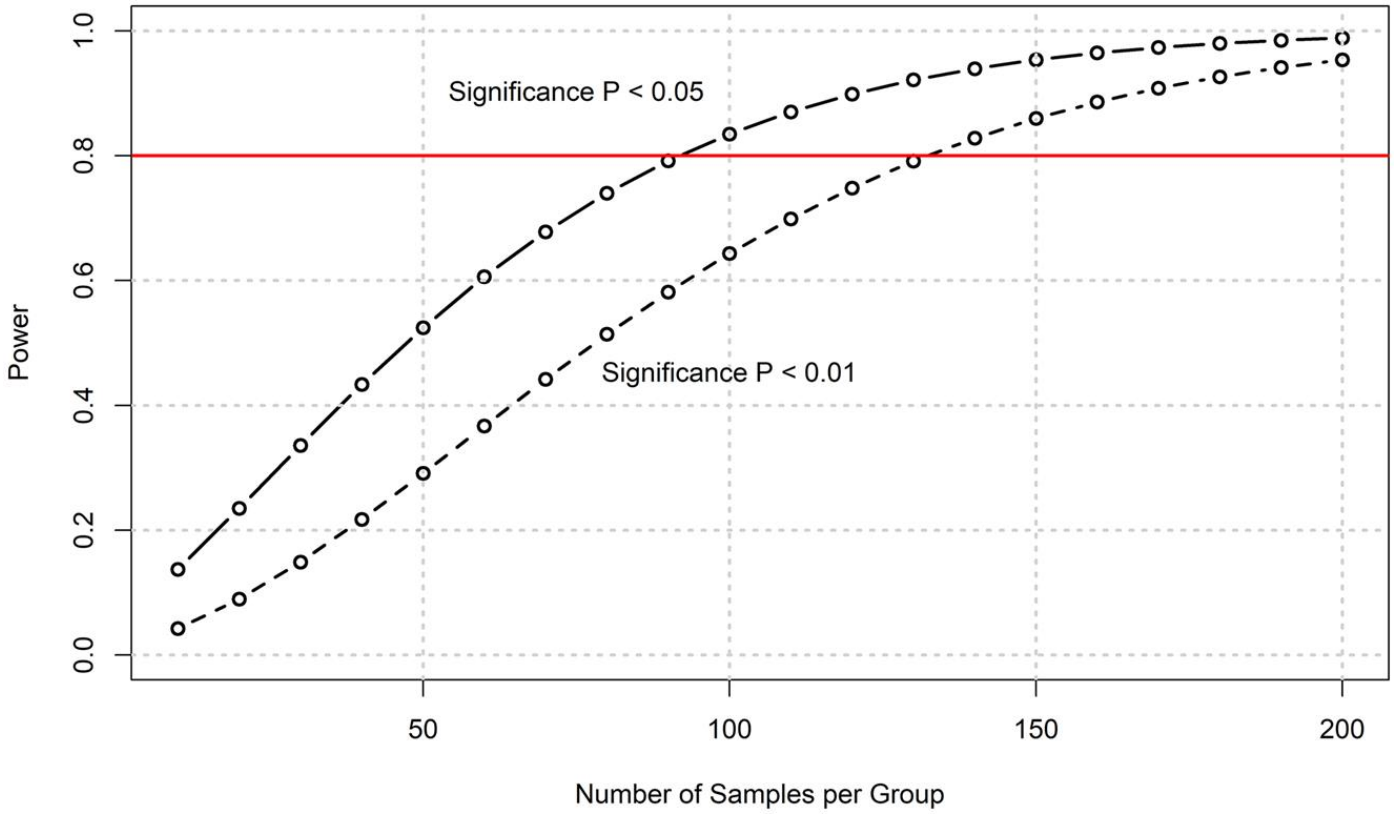


The x-axis represents the 'Number of Monkeys per Group' ranging from 0 to 20 in increments of 5. The y-axis represents 'Power' ranging from 0 to 1 in increments of 0.2. Two lines are plotted on the graph: 1. A solid line representing 'Significance P less than 0.05' 2. A dashed line representing 'Significance P less than 0.01'. Both lines show an increasing trend as the number of monkeys per group increases. A horizontal red line is drawn at the 0.8 power level, which both curves cross between 5 and 10 monkeys per group. The detailed data points are as follows: For 1 monkey per group, the power with P less than 0.05 is 0.17 and the power with P less than 0.01 is 0.05. For 5 monkeys per group, the power with P less than 0.05 is 0.47 and the power with P less than 0.01 is 0.15. For 10 monkeys per group, the power with P less than 0.05 is 0.82 and the power with P less than 0.01 is 0.65. For 15 monkeys per group, the power with P less than 0.05 is 0.95 and the power

with P less than 0.01 is 0.85. For 20 monkeys per group, the power with P less than 0.05 is 0.98 and the power with P less than 0.01 is 0.93.

Navigate back to Figure 5.4

Description of Figure 5.5

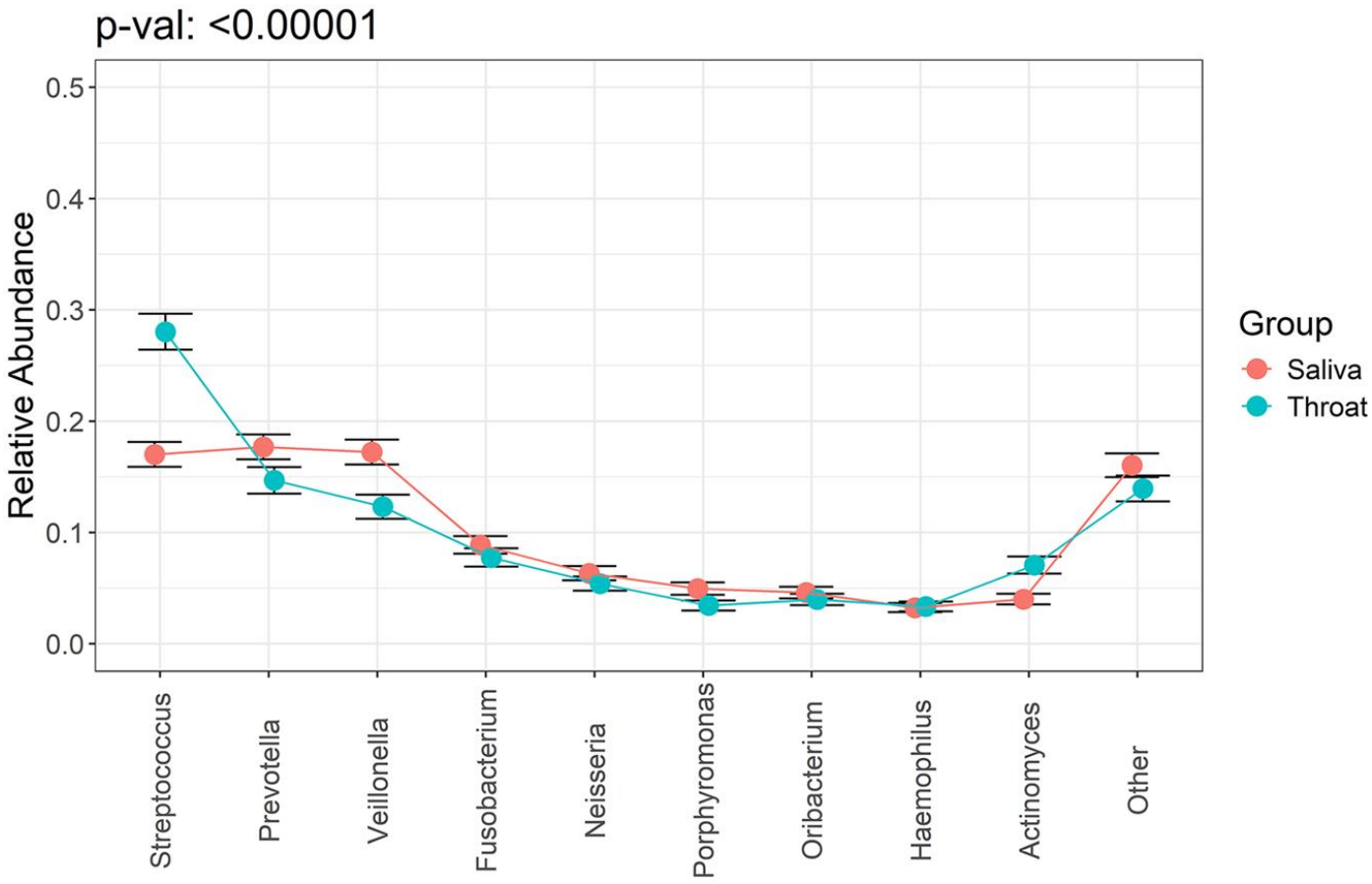


The x-axis represents the 'Number of Monkeys per Group' ranging from 0 to 20 in increments of 5. The y-axis represents 'Power' ranging from 0 to 1 in increments of 0.2. Two lines are plotted on the graph: 1. A solid line representing 'Significance P less than 0.05' 2. A dashed line representing 'Significance P less than 0.01'. Both lines show an increasing trend as the number of monkeys per group increases. A horizontal red line is drawn at the 0.8 power level, which both curves cross between 5 and 10 monkeys per group. The detailed data points are as follows: For 1 monkey per group, the power with P less than 0.05 is 0.17 and the power with P less than 0.01 is 0.05. For 5 monkeys per group, the power with P less than 0.05 is 0.47 and the power with P less than 0.01 is 0.15. For 10 monkeys per group, the power with P less than 0.05 is 0.82 and the power with P less than 0.01 is 0.65. For 15 monkeys per group, the power with P less than 0.05 is 0.95 and the power

with P less than 0.01 is 0.85. For 20 monkeys per group, the power with P less than 0.05 is 0.98 and the power with P less than 0.01 is 0.93.

Navigate back to Figure 5.5

Description of Figure 5.6



The x-axis represents the 'Number of Samples per Group' ranging from 0 to 200 in increments of 50. The y-axis represents 'Power' ranging from 0 to 1.0 in increments of 0.2. Two curves are plotted on the graph: 1. A curve labeled 'Significance P less than 0.05' 2. A curve labeled 'Significance P less than 0.01'. Both curves show an increasing trend as the number of samples increases. A horizontal red line is drawn at the 0.8 power level, which both curves cross. The detailed data points are as follows: For 25 samples per group, the power for P less than 0.05 is 0.23 and the power for P less than 0.01 is 0.08. For 50 samples per group, the power for P less than 0.05 is 0.52 and the power for P less than 0.01 is 0.29. For 75 samples per group, the power for P less than 0.05 is 0.74 and the power for P less than 0.01 is 0.51. For 100 samples per group, the power for P less than 0.05 is 0.86 and the

power for P less than 0.01 is 0.71. For 125 samples per group, the power for P less than 0.05 is 0.93 and the power for P less than 0.01 is 0.83. For 150 samples per group, the power for P less than 0.05 is 0.96 and the power for P less than 0.01 is 0.90. For 175 samples per group, the power for P less than 0.05 is 0.98 and the power for P less than 0.01 is 0.94. For 200 samples per group, the power for P less than 0.05 is 0.99 and the power for P less than 0.01 is 0.97.

Navigate back to Figure 5.6

CHAPTER 6

An update and review of ribosomal RNA depletion methods for microbial transcriptomics

Shaomei He, Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, United States

Abstract

Over the last two decades the rapid advancement of high-throughput sequencing technologies has revolutionized the field of microbial transcriptomics, allowing large-scale detection and quantification of gene expression in single cells, microbial populations, and complex communities. These advancements have enabled unprecedented discoveries and understanding of microbial gene function and regulation. Despite this one major challenge in sequence-based microbial transcriptomics is the overwhelming abundance of ribosomal RNA (rRNA) in total RNA, which limits the detection and quantification of more informative messenger RNA (mRNA) and regulatory RNAs, thus requiring higher sequencing depths for detecting rare transcripts. To overcome this obstacle a variety of rRNA depletion methods have been developed, each with its own strengths and limitations. This chapter aims to review rRNA depletion methods established over the years, highlighting their advantages, limitations, and potential biases in the context of microbial transcriptomics.

Keywords

Microbial transcriptomics; metatranscriptomics; RNA-seq; ribosomal RNA removal; messenger RNA enrichment; subtractive hybridization; RNase H digestion; CRISPR/Cas9; cDNA library normalization; not-so-random primer; single-cell RNA-seq; dual RNA-seq; total RNA-seq

Introduction

Transcriptomics studies the transcriptome, which is the complete set of transcribed sequences in a cell or a population of cells, including messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), and regulatory noncoding RNA (ncRNA). Metatranscriptomics, also sometimes referred to as "environmental transcriptomics" or "microbial community transcriptomics," studies the entire collection of transcribed sequences from microbial communities (Moran, 2009). As a powerful tool in the postgenomics era transcriptomics and metatranscriptomics (collectively referred to as microbial transcriptomics in this chapter) facilitate the understanding of gene structure and regulation, provide a snapshot of gene expression under specific conditions, and reveal functional regulation and interactions responding to environmental changes (Filiatrault, 2011; Sorek & Cossart, 2010; Moran, 2009).

Microbial transcriptomics is evolving with technological advancement, from constructing cDNA clone libraries, to oligonucleotide microarrays, and to high-throughput sequencing technology-enabled transcriptome shotgun sequencing (RNA-seq). With the increase in sequencing throughput and decrease in cost over the years RNA-seq has become a popular and powerful tool in microbial transcriptomics.

However a major technical challenge for RNA-seq is the low relative abundance of mRNA in total cellular RNA (1%–5%) (Neidhardt & Umbarger, 1996), the bulk of which consists of rRNA and tRNA (Karpinets et al., 2006), particularly rRNA. For example rRNA sequences often constituted more than 90% of the total reads in RNA-seq (Giannoukos et al., 2012; He et al., 2010; Peano et al., 2013; Yi et al., 2011). This rRNA predominance wastes enormous

 amounts of sequencing reads, limits the detection of rare genes and populations, and hinders statistical comparison of gene expression. Therefore researchers often chose to eliminate rRNA before sequencing to improve mRNA detection sensitivity. Unlike eukaryotic mRNAs, which have poly(A) tails (Zhao et al., 1999) and can be selectively synthesized into cDNA using poly(T) primers, prokaryotic mRNAs lack poly(A) tails long enough for this application. This had greatly hindered RNA-seq studies on prokaryotes.

Over the years various methods have been developed to eliminate prokaryotic rRNA with varying degrees of success. These methods leverage unique characteristics of rRNA relative to non-rRNA, such as size, specific sequences, 5′-phosphate, secondary structure, and abundance, to physically separate rRNA from non-rRNA, enzymatically degrade rRNA, or preferentially synthesize or amplify from non-rRNA (**Table 6.1**). The depletion step can occur before, during, or after cDNA synthesis, either with or without RNA amplification, thereby having different requirements for total RNA input quantities (e.g., depletion rRNA before cDNA synthesis without any amplification typically demands higher input RNA quantities) (**Table 6.1**). Among these methods those utilizing rRNA-specific sequences constitute the largest category and have been widely applied in RNA-seq studies, particularly through various commercial kits.

Table 6.1

Method		rRNA feature employed	RNA input	Commercial kit
Prior to cDNA synthesis				
Size selection by gel electrophoresis		Size (and high abundance)	High	
Denaturing high-performance liquid chromatography		Hydrophobicity (influenced by rRNA size and sequence composition)	High	
5′ phosphate-dependent exonuclease digestion		5′-monophosphate of processed RNA	High	Epicenter mRNA-ONLY prokaryotic mRNA isolation kit
Subtractive hybridization	Generic probes	Sequences (for generating probes/primers complementary to rRNA)	Medium to high	Invitrogen's MICROBExpress bacterial mRNA enrichment kit; Illumina Ribo-Zero rRNA depletion kits (Bacteria) (discontinued); Thermo Fisher's RiboMinus transcriptome isolation kit; siTOOLs' riboPOOLs rRNA depletion kit (Single-Species, Pan-Prokaryote, Pan-Bacteria, Pan-Archaea); Lexogen's RiboCop rRNA depletion kit for Bacteria
	Custom probes			siTOOLs' CustomriboPOOLs rRNA depletion kit
RNase H digestion	Generic probes			New England Biolabs' NEBNext rRNA depletion kit; KAPA RiboErase; Illumina's Ribo-Zero Plus rRNA depletion kit (microbiome); Vazyme's Ribo-off rRNA Depletion Kit V2 (bacteria)
	Custom probes			New England Biolabs' NEBNext Depletion Core Reagent Set with customized sequence-specific probes
Biased 3' polyadenylation against rRNA	Without amplification	Complex with ribosome and form secondary structure	High	NEBNext Poly(A) mRNA Magnetic Isolation Module
	With IVT amplification		Low	Thermo Fisher's MessageAmp-II bacteria RNA amplification kit
During cDNA synthesis				
Blocking reverse transcription of rRNA in library preparation	Generic or custom blocking oligos	Sequences (for generating blocking oligos complementary to rRNA)	Low to Medium	QIAGEN's QIAseq FastSelect 5S/16S/23S kit
Selective priming with not-so-random oligomers	Without amplification	Differential occurrence of hexamers in rRNA relative to non-rRNA	Medium	
	With PCR amplification		Low	NuGEN's Ovation RNA-Seq system; Tecan's Universal Prokaryotic RNA-Seq, Prokaryotic AnyDeplete (selective primers are customizable)
Selective amplification of mRNA by hybridization with mRNA-specific probes	With IVT amplification	Sequence difference with mRNA genome-wide	Very low	
After cDNA synthesis				
Post hoc RNase H digestion	With IVT or PCR + IVT amplification	Sequences (for generating probes, sgRNA, or blocking primers complementary to rRNA)	Very low	Same RNaseH digestion kitslisted above
CRISPR/Cas9 digestion	Generic guide RNA		Low	Jumpcode Genomics' CRISPRclean Plus Stranded Total RNA Prep with rRNA Depletion (Human, Mouse, Rat, Pan-Bacteria) kit
	Custom guideRNA		Low	Jumpcode Genomics' CRISPRclean reagents with custom guide RNA design
Post hoc CRISPR/Cas9 digestion	With PCR, IVT, or random displacement amplification		Very low	Same CRISPR/Cas9 digestion kits listedabove

Method		rRNA feature employed	RNA input	Commercial kit
Enrichment of mRNAby blocked rRNA (EMBR)	With IVT amplification		Very low	
cDNA library normalization	With duplex-specific nuclease digestion	High abundance of rRNA	Low	Evrogen's Trimmer-Direct cDNA normalization kit; Zymo Research's Zymo-Seq RiboFree total RNA library kit
	With hydroxyapatite chromatography separation			

IVT, *in vitro* transcription; *PCR*, polymerase chain reaction; *CRISPR*, clustered regularly interspersed short palindromic repeats; *Cas9*, CRISPR-associated protein 9.

So, which method should be chosen? In addition to the input RNA quantity, which may constrain options, the two most critical considerations in method selection are rRNA depletion efficiency and potential bias. The quality of input RNA (e.g., intact vs fragmented or degraded) and the specific microbial species present in the sample can significantly impact efficiency and bias. While some limitations and biases are readily recognized, others may be less obvious. Numerous studies have systematically evaluated and compared rRNA depletion methods and kits (Choe et al., 2021; Ciulla et al., 2010; Culviner et al., 2020; Giannoukos et al., 2012; He et al., 2010; Heom et al., 2023; Huang et al., 2020; Kim et al., 2019; Pastor et al., 2022; Peano et al., 2013; Petrova et al., 2017; Wang et al., 2023; Xiang et al., 2022; Yi et al., 2011; Bhagwat et al., 2014; Engelhardt et al., 2020; Stewart et al., 2010; Wahl et al., 2022). In these studies efficiency is measured by the proportion of rRNA reads in the total RNA sequenced with and without rRNA depletion. Bias is primarily assessed through fidelity, which refers to the ability to preserve the relative abundance of mRNA transcripts after rRNA depletion compared with the original crude Page 111 Page 112 Page 113 extract. The fidelity of the depletion procedure is often evaluated with the correlation coefficient (r) or coefficient of determination (R^2) between mRNA transcript abundance with

and without rRNA depletion. Given these insights and advancements made over the last two decades this chapter provides an overview of the various methodologies for rRNA depletion, discussing their advantages, limitations, and potential biases. These methods are presented in an order primarily based on the rRNA features utilized and the timing of rRNA depletion (i.e., before, during, or after cDNA synthesis) (**Table 6.1**) in the following sections. With the diversity of studies and their associated resource availability, none of these techniques are "off the table" for present and future applications.

Size selection by gel electrophoresis

Due to their extremely high abundance in the total RNA and their distinct sizes, intact 5, 16, and 23 S rRNAs are easily recognizable as distinct sharp bands on an agarose gel after the total RNA is subjected to electrophoresis. Agarose gel electrophoresis separates charged molecules (such as RNA) based on their sizes/lengths, as the speed of these molecules to migrate through the gel in the electric field is inversely related to their molecular sizes/lengths. To recover mRNA after gel electrophoresis McGrath et al. (2008) extracted RNA from regions between the 23 and 16 S bands on the agarose gel for long mRNAs, and between the 16 and 5 S rRNA bands for short mRNAs, thereby excluding 5, 16 and 23 S rRNAs. They tested this method on *Escherichia coli* and a number of environmental samples and observed increased recovery of mRNA in the resulted cDNA clone library (McGrath et al., 2008). A similar approach was also applied to *E. coli*, where the gel extraction for long and short mRNA regions was conducted after the tRNA removal by a size exclusion chromatography (Petrov et al., 2022). The limitations of these methods are obvious. First fragmented rRNA or pre-rRNA (precursor molecules before being cleaved into 5, 16 and 23 S rRNAs) not at the expected sizes of intact mature rRNA cannot be removed, while mRNA that comigrates at the same speed as rRNA, due to similarity in size and electrophoretic behavior, will not be recovered. In addition the amount of

input RNA needs to be high enough for gel visualization and subsequent mRNA recovery [e.g., 10 µg of RNA in the above study by Petrov et al. (Petrov et al., 2022)], and it is more challenging to maintain an RNase-free Page 114 environment in the gel, buffer, and electrophoresis equipment than in a single tube.

Denaturing high-performance liquid chromatography

The physical separation of RNA can also be achieved through RNA chromatography under full denaturing conditions (75°C) that thermally disrupt the secondary structure of RNA molecules. In this method the stationary phase consists of a nonporous alkylated poly(styrene-divinylbenzene) bead matrix that provides a highly hydrophobic surface (Azarani & Hecker, 2000). The separation primarily relies on the strength of the hydrophobic interaction between RNA and the bead matrix, which is largely determined by the size and base composition (sequence) of the denatured RNA. The hydrophobicity-based separation and selective collection of RNA are performed by applying a gradient of acetonitrile (an organic solvent in the mobile phase), with gradual increase in acetonitrile concentration to elute RNA with increasing hydrophobicity from the bead matrix. Castro et al. (2013) adopted denaturing high-performance liquid chromatography for RNA separation and removed 78%–92% of rRNA extracted from *Corynebacterium pseudobuberculosis*. This method provides rapid separation of bulk rRNA and mRNA, completed within half an hour. However, it requires a high RNA input [e.g., 20 µg (Castro et al., 2013)] to effectively recover mRNA. In addition small RNAs elute faster and can be missed from the recovery. Whether the chromatographic profile-based RNA separation is applicable for other microorganisms or for complex environmental samples remains to be examined.

5 ′ phosphate-dependent exonuclease digestion

Terminator 5´-Phosphate-Dependent Exonuclease, a processive 5′ to 3′ exonuclease, selectively degrades RNA molecules with 5′-monophosphates, thereby leaving RNAs with 5′-triphosphate or hydroxyl groups intact. As mature 5, 16 and 23 S rRNAs are processed from a single pre-rRNA transcript, they possess 5′-monophosphates making them susceptible to the degradation by this enzyme. In contrast, intact mRNAs carry 5′-triphosphates that are resistant to the degradation by this enzyme. This rRNA depletion strategy, based on the general feature at the 5' end of processed rRNAs, is expected to be broadly effective for bacteria and archaea, making it potentially useful for complex microbial communities. However unexpected low removal Page 115 efficiency of rRNAs from an archaeal *Halorhabdus* sp. by this approach was observed (He et al., 2010). The authors suspected that some of the processed 5′ ends of archaeal rRNAs may have an exposed hydroxyl group rather than a monophosphate, as also suggested by the poor ligation of archaeal *Sulfolobussolfataricus* rRNAs to adapters that require an exposed 5′-monophosphate (Wurtzel et al., 2010). The maturation of rRNA in archaea is not well characterized and may deviate from the *E. coli* pathway. This also underlines the importance of testing kits when working with phyla that have not been included in the kit development.

The effectiveness and fidelity of this method can be affected by the presence of partially degraded mRNA in crude extracts (He et al., 2010). In bacteria and archaea many mRNAs have short half-lives and can degrade in minutes after transcription (Rauhut & Klug, 1999). Continuous transcription is needed to maintain translation. Due to the rapid turnover many extracted mRNAs may be partially degraded. One mechanism of mRNA degradation in bacteria starts with the conversion of 5′-triphosphates to monophosphates, which facilitates subsequent internal cleavage of mRNA by RNase E (Jiang & Belasco, 2004). Consequently a fraction of mRNAs is monophosphorylated and fragmented and thus susceptible to the 5′-monophosphate-specific exonuclease digestion. Indeed mRNAs with short half-lives and high turnover rates, presumably partially degraded, were preferentially lost after 5′-

monophosphate-specific exonuclease digestion (He et al., 2010). Such unintended mRNA degradation not only compromises mRNA fidelity but also counteracts the effectiveness of rRNA depletion (He et al., 2010). Conceivably this enzymatic treatment could provide a useful snapshot of stable full-length mRNA, and theoretically it would be useful to identify transcription start sites due to its enrichment for primary transcripts with 5′-triphosphates.

Subtractive hybridization

The most prevalent category of rRNA depletion methods utilizes sequences complementary to rRNA for its removal. These sequences can serve as capture probes for the physical separation of rRNA from mRNA in subtractive hybridization, as primers/probes and guide RNA for selective enzymatic digestion in RNase H- and CRISPR/Cas9-based methods, respectively, or as blocking oligos or blocking primers to inhibit cDNA synthesis from rRNA or to prevent sequence Page 116 amplification via *in vitro* transcription from rRNA-derived cDNA, respectively (**Table 6.1**). Among these approaches subtractive hybridization takes place before cDNA synthesis; RNase H digestion can occur either before or after cDNA synthesis; blocking oligos and primers are used during and after cDNA synthesis, respectively; while the CRISPR/Cas9-based digestion is performed after cDNA synthesis (**Table 6.1**). The subtractive hybridization approach is discussed in this section, and the other antisense rRNA sequence-based approaches will be discussed in the following four sections.

In subtractive hybridization probes are first hybridized to rRNA. As the probes are labeled, the resulting rRNA:probe hybrids can bind to correspondingly coated magnetic beads, allowing for the subtraction of rRNA from solution (Pang et al., 2004). This method has been the most widely used to date, partly owing to the availabilities of various commercial kits (**Table 6.1**).

The MICROBExpress Bacterial mRNA Enrichment kit from Invitrogen is one of the earliest commercial kits designed for depleting rRNA from

bacterial samples. Its capture probes are short DNA oligonucleotides with one end targeting conserved regions of 16 and 23 S rRNAs and the other end containing a complementary sequence to a secondary probe bound to magnetic beads. This probe configuration allows the capture of rRNA:probe hybrids to the beads after two sequential hybridizations. A major strength of this kit is its high fidelity in rRNA removal, as studies have shown that the correlations of mRNA transcript abundance before and after rRNA depletion were usually higher with MICROBExpress compared with other evaluated methods (Giannoukos et al., 2012; He et al., 2010; Peano et al., 2013). For example, only a small number (2%–3%) of transcripts exhibited greater than twofold change in expression relative to the sample without rRNA removal, suggesting that probe cross-hybridization to mRNA was minimal (He et al., 2010).

However the removal efficiency is highly dependent on organism(s) present in the sample, since the capture probes are only compatible with a small subset of bacteria while the entire archaeal domain is not targeted. In some applications two rounds of MICROBExpress were applied sequentially on metatranscriptome samples (Shrestha et al., 2009), presumably to improve mRNA enrichment. However, upon evaluation, two rounds did not produce a significant improvement in rRNA depletion over one round, suggesting that rRNAs with target sites complementary to the capture probes were efficiently removed in Page 117 the first round and the additional round is unnecessary (He et al., 2010). In addition to the limited target range of the capture probes the success of this kit is also affected by RNA integrity, as demonstrated by the inverse correlation between rRNA removal and RNA integrity (He et al., 2010). As a very small number of capture probes are employed by this kit, RNA degradation and fragmentation can increase the proportion of rRNA fragments lacking the targeting sites complementary to the conserved capture probes, and this can compromise the rRNA removal efficiency for samples with low RNA integrity.

The later-introduced Ribo-Zero rRNA Removal Kit (Bacteria) from Illumina was the most widely used rRNA removal kit for prokaryotic RNA-seq before its discontinuation in 2018. This kit was based on the Patent US20110040081A1 (Sooknanan et al., 2010) and utilized biotinylated antisense rRNA as probes that can be subsequently captured by streptavidin-coated magnetic beads. Compared with MICROBExpress, Ribo-Zero removed not only 16 and 23 S rRNA but also 5 S rRNA, its probes targeted a broader range of Gram-positive and Gram-negative bacteria. Notably a small number of archaeal species were also included, and for this reason, Ribo-Zero kit was frequently employed in archaeal RNA-seq studies (Pastor et al., 2022). In addition it used antisense rRNA probes that nearly cover the entire lengths of rRNAs. Due to these improvements Ribo-Zero had significantly higher depletion efficiency than MICROBExpress while maintaining a high fidelity in numerous systematic comparisons (Ciulla et al., 2010; Peano et al., 2013; Bhagwat et al., 2014; Wahl et al., 2022). In addition Ribo-Zero is also effective on RNA of low integrity, as confirmed by its performance on a highly fragmented RNA sample (Giannoukos et al., 2012). However its application on low-input RNA (e.g., below microgram quantities) should be cautioned, as the effectiveness and fidelity were greatly compromised in such cases (Giannoukos et al., 2012).

Other available subtractive hybridization-based commercial kits later introduced into the market include the RiboMinus Transcriptome Isolation Kit (Bacteria) from Thermo Fisher, riboPOOLs rRNA Depletion Kits from siTOOLs Biotech, and RiboCop rRNA Depletion Kit for Bacteria from Lexogen (**Table 6.1**). As subtractive hybridization is applied before cDNA synthesis, these kits often require relatively high total RNA inputs. For example, the earlier introduced subtractive hybridization kits require 1–5 μg of total RNA, although Page 118 the later introduced riboPOOLs and RiboCop have reduced the minimal RNA input to as low as 10 ng.

In comparative studies where MICROBExpress, Ribo-Zero, and RiboMinus were compared on bacterial isolates or cocultures, Ribo-Zero consistently

yielded the highest degree of rRNA depletion with the least bias (Petrova et al., 2017; Bhagwat et al., 2014). Due to its superior performance, Ribo-Zero was regarded as a gold standard for rRNA depletion and was widely applied in microbial RNA-seq before its discontinuation in 2018.

The later-introduced riboPOOLs kit has been suggested as an effective alternative to Ribo-Zero in a more recent comparative study where four commercial subtractive hybridization kits along with an in-house method were tested on *E. coli* (Wahl et al., 2022). Among these kits riboPOOLs exhibited performance comparable to Ribo-Zero, both achieving much higher depletion efficiencies than RiboMinus and MICROBExpress. In a separate study where riboPOOLs and Ribo-Zero were applied to *Salmonella typhimurium*, the former kit reduced rRNA reads to ~3% and the latter to 0.1%, and they showed a high correlation (Pearson $r = 0.98$) in their overall transcript abundance (Kim et al., 2019).

Notably riboPOOLs offer different versions, such as single-species riboPOOLs (designed to target specific organisms), Pan-riboPOOLs (including Pan-Prokaryote, Pan-Bacteria, and Pan-Archaea, targeting a wide range of diverse organisms), and Custom riboPOOLs (a customizable option that allows users to design and synthesize custom probes from specified sequences). While most commercial kits are designed primarily for bacteria, riboPOOLs is a great option for archaeal RNA-seq, as its Pan-Archaea probes match more than 350 archaeal species. Indeed a comparative analysis of rRNA removal methods for haloarchaea found that riboPOOLs Pan-Archaea with its generic probes efficiently removed rRNA across four haloarchaea species and outperformed Ribo-Zero in some cases (Pastor et al., 2022). Furthermore its custom-designed probes even resulted in nearly complete rRNA depletion for the targeted haloarchaea species (Pastor et al., 2022). However, unlike Ribo-Zero whose probes nearly cover the entire lengths of rRNAs, the Pan-riboPOOLs probes were designed to target only a number of regions that are conserved among the large number of species for even coverage across different taxa. Therefore, for effective depletion, the

manufacturer recommends using input RNA with relatively high integrity as indicated by the value of ≥7 for Page 119 the RNA integrity number (RIN), which is calculated using several characteristics of an RNA microcapillary electropherogram, with values ranging from 1 (most degraded) to 10 (least degraded).

Although commercial kits are convenient and typically achieve high transcript fidelity, they can be expensive and are often limited by the target range offered by generic probes. In contrast do-it-yourself (DIY) kits provide a cost-effective, customizable alternative with probes tailored for individual organisms or microbial communities, enabling sample-specific subtraction. In fact, before commercial kits became available, early versions of subtractive hybridization were organism-specific, using biotin-labeled full-length or near full-length rRNA genes cloned into a plasmid (Robinson et al., 1994) or antisense rRNA *in vitro* transcribed from cloned rRNA genes (Plum & Clark-Curtiss, 1994; Su & Sordillo, 1998). Later Li et al. (2013) developed an organism-specific subtraction system. They first designed a computer program to select probes specific to 16 and 23 S rRNA of the target organism without cross-hybridization to its mRNAs. The probe sequences were then PCR-amplified from cDNA, cloned into a plasmid system, and *in vitro* transcribed to generate biotin-labeled antisense rRNA probes. They tested this system on *Mycobacterium smegmatis* and obtained significant rRNA reduction with minimal impact on mRNA.

Similarly Stewart et al. (2010) published a microbial community-specific rRNA depletion method. This method starts with extracting community DNA and RNA in parallel from the same sample. Near full-length 16 S and 85% of the 23 S rRNA genes are PCR-amplified from the community DNA using universal primers containing the T7 RNA polymerase promoter sequence at the 5′ end of the reverse primers, allowing for subsequent *in vitro* transcription and the generation of biotin-labeled antisense rRNAs complementary to the community rRNA. Ideally probes generated from community RNA provide a more accurate representation of rRNA

composition than those derived from DNA, as different community members may have different levels of activity and rRNA contents, which are not accurately reflected by the rRNA gene abundance. However reverse transcription is less efficient for long transcripts (e.g., 23 S rRNA), as the low processivity of reverse transcriptase causes it to dissociate from the template more easily than high-fidelity DNA polymerase, and the complex secondary structures of long rRNA further increase the enzyme's tendency to fall off the template (Lucas et al., 2023). Therefore using DNA as templates is more efficient for synthesizing long antisense rRNA probes to maximize the coverage of rRNA molecules. When applied on marine metatranscriptome samples, Stewart et al. (2010)successfully reduced rRNA to an average of 36% of total reads.

As a single primer set may not amplify all phylogenetic groups in a complex microbial community, and the secondary structure of near full-length rRNAs may decrease hybridization efficiency, Kukutla et al. (2013) generated probes from four combinations of two forward and two reverse primers. They pooled these probes to maximize phylogenetic coverage and tested them on four mosquito gut metatranscriptome samples. Some primer combinations generated shorter probes, which presumably are less prone to forming secondary structure, and thus may enhance the probe-rRNA hybridization. Following the depletion procedure and sequencing, the rRNA reads accounted for only 1%–11% of the total reads for the four samples tested, indicating high removal efficiencies (Kukutla et al., 2013).

More recently improved DIY kits with customizable probes have been developed, featuring more sophisticated probe design strategies to improve effectiveness while reducing cross-hybridization to mRNA (Culviner et al., 2020; Kim et al., 2019; Wahl et al., 2022). These self-made depletion protocols utilize biotinylated probes and streptavidin-coated magnetic beads but differ in the probe design and synthesis approaches. For instance Culviner et al. (2020) designed and chemically synthesized large pools of biotinylated short DNA oligonucleotides (40 and 15–24 nt in length). In contrast probes used by

Wahl et al. (2022) were near full-length antisense rRNA synthesized in-house from gDNA via PCR with rRNA-specific primers, followed by *in vitro* transcription. This method results in RNA:RNA hybrids, which are more stable than RNA:DNA hybrids, enabling faster hybridization and more efficient removal (Wahl et al., 2022). This probe synthesis method resembles the approach used by Stewart et al. (2010), with one improvement being that the long 23 S rRNA gene is amplified in two separate PCRs with two sets of primers, each set targeting one half of the gene to achieve more effective coverage of its full length.

Published in-house methods were compared against the gold standard Ribo-Zero and achieved comparable efficiency with minimal bias on tested bacterial isolates (Culviner et al., 2020; Kim et al., 2019; Wahl et al., 2022), therefore offering promising customizable alternatives. However, even with custom probe design, rRNA removal can still be challenging for complex environmental communities where diverse Page 121 species with varying activities coexist at different abundance levels. It is expensive to chemically synthesize the large number of probes and combine them at the right ratios to effectively target the entire community. Therefore the protocol developed by Wahl et al. (2022) using in-house synthesized, sample-specific antisense rRNA probes, although only tested on *E. coli* , may be a viable choice for complex communities, by carefully designing additional primers to broaden the target taxonomic range.

A challenge with homemade methods is that they often require expensive and laborious reoptimization through trial-and-error when applied to new samples or when reagents or protocols need to be changed. To address this a systematic framework, the statistical design of experiments (DOEs), was proposed for such reoptimization. DOE uses systematically controlled tests to identify factors and variables with significant impacts on the desired output and to understand the interactions among variables and their combined effects on the output for efficient optimization and prediction of the output. DOE was applied to improve a subtractive hybridization-based rRNA

depletion protocol by exploring factors including total RNA input quantity, probe concentration, percent of biotin-labeled UTP, and type and volume of streptavidin magnetic bead input; protocol settings for these factors were further optimized to maximize rRNA removal while minimizing the cost (David & Jensen, 2023). The framework proposed in this study would be valuable for large-scale transcriptomics and metatranscriptomics studies.

RNase H digestion

Another rRNA depletion method that employs rRNA-targeted probes (or primers) relies on enzymatic rRNA elimination using RNase H (Hausen & Stein, 1970) that specifically degrades the RNA in DNA:RNA hybrids, leaving single-stranded RNA intact. In an early study published in 2001 Dunman et al. (2001) applied RNase H-mediated rRNA depletion to profile *Staphylococcus aureus* gene expression. Total RNA was first subjected to reverse transcription with a mix of primers specific to 16 and 23 S rRNAs. Subsequently RNase H was added to degrade the RNA fraction of the cDNA:RNA duplex, followed by a DNase I treatment to eliminate the cDNA fraction. Therefore optimization of these enzymatic reactions for complete RNase H and DNase I digestions is required for the effectiveness Page 122 and fidelity of this method. Additionally the specificity and taxonomic coverage of primer sequences, along with the annealing conditions used during reverse transcription, can significantly impact its performance. Furthermore fragmented rRNA lacking primer binding sites cannot be removed. Moreover, due to the low processivity of reverse transcriptases and the complex secondary structures of rRNA, the reverse transcriptase can fall off templates. As a result longer transcripts are less likely to be fully reverse transcribed, compromising the removal of the 5′ ends of rRNA, especially 23 S rRNA. Multiple primers targeting multiple sites across rRNA molecules are needed to improve effectiveness of this approach.

Due to these limitations, the RNase H-based method did not gain much popularity until the discontinuation of the most widely used Ribo-Zero

subtractive hybridization kits in 2018. In 2020 Huang et al. (2020) developed a cost-effective and customizable RNase H method, which addressed some of the above limitations. In this approach DNA:RNA hybrids are generated by hybridization with single-stranded DNA (ssDNA) probes, rather than through reverse transcription as used by Dunman et al. (2001). The ssDNA probes can be synthesized chemically as ~50 nt oligonucleotides covering the entire lengths of 16 and 23 S rRNA or produced by PCR amplification from gDNA using universal primers, which is more efficient than reverse transcription for covering long RNAs. For the oligo-based probes an online tool was developed to design probe libraries, allowing for customizable applications. For the PCR amplicon-based probes the forward primer is 5′-phosphorylated, allowing lambda exonuclease to selectively digest the 5′-phosphorylated strand of the double-stranded DNA amplicon to generate ssDNA probes, enabling a sample-specific method for rRNA depletion. This approach is compatible with barcoded or fragmented RNA and achieved high efficiencies (comparable to Ribo-Zero) and fidelity (R^2>0.94) for tested bacterial species from three distinct phyla (Huang et al., 2020). Later the same year Engelhardt et al. (2020) also published an RNase H method that employs probes of 50±5 nt in length across the entire lengths of unwanted RNA species (including 16 and 23 S rRNA, and tmRNA) in *Pseudomonas aeruginosa*. This method achieved greater depletion than two tested commercial kits, including the Ribo-Zero (subtractive hybridization-based) and the Ribo-off kit (RNase H-based). These two studies present a promising customizable approach for sample-specific removal of unwanted RNA species by RNase H treatment. Page 123

An optimized RNase H method, termed RiboRid, was further developed to minimize nonspecific hybridization to mRNA and to enhance rRNA depletion efficiency (Choe et al., 2021). Probes were designed to target the entire lengths of 16 and 23 S rRNA, as well as pre-rRNA. To ensure hybridization efficiency and specificity, a Python script considering melting temperatures was used for probe design. Additional optimizations included adjusting the reaction temperature and ionic strength. When tested on seven diverse

bacterial species, the optimized method attained high rRNA removal efficiencies (~99%) and fidelity (R^2>0.91) in most cases, which were comparable to the performance of Ribo-Zero, but at a significantly lower cost.

Several companies have commercialized the RNase H-based approach, offering kits such as the NEBNext rRNA Depletion Kit (bacteria) from New England Biolabs, the Ribo-off rRNA Depletion Kit (bacteria) from Vazyme, and the Ribo-Zero Plus Microbiome Depletion Kit from Illumina. Notably this RNase H-based Ribo-Zero Plus Microbiome Depletion Kit is very different from the original subtractive hybridization-based Ribo-Zero kit from the same manufacturer. These kits utilize ssDNA probes that span the entire lengths of 5, 16, and 23 S rRNA, making them effective on even fragmented and degraded RNA samples. They all involve hybridization with probes, RNase H digestion, DNase I digestion, and cleanup using magnetic beads.

Similar to subtractive hybridization the efficiency of RNase H-based methods depends on probe specificity and coverage and thus can be limited by the nonspecific generic probes included in the commercial kits. For instance, using the NEBNext rRNA Depletion Kit (bacteria), with its generic probes targeting bacteria, resulted in low removal efficiencies for archaea like haloarchaea (Pastor et al., 2022). To address this limitation some commercial kits offer options for using custom probes. The NEBNext® Custom RNA Depletion Design Tool (**https://depletion-design.neb.com**) allows users to design probes targeting taxa-specific sequences to be used together with the NEBNext RNA Depletion Core Reagent set. In the abovementioned study with haloarchaea, when custom probes designed for a specific haloarchaea species were used in the NEBNext kit, near-complete rRNA removal was achieved for the specific species (Pastor et al., 2022). Similarly the Ribo-Zero Plus workflow also offers such an option. For example the standard probes in the Ribo-Zero Plus kit were not efficient in depleting rRNA from human microbiome samples, and thus an iterative design process was implemented to generate additional custom Page 124 probes to combine with the standard generic probes to enhance rRNA depletion efficiency. When

applying Ribo-Zero Plus on stool samples from 10 donors, the average percentage of bacterial rRNA reads in total reads was 70% with the standard generic probes and decreased to 17% with a mix of standard and custom probes; and on oral samples this percentage decreased from 45% to 5% by adding custom probes, highlighting the benefits of including custom probes (Tan et al., 2023).

In general subtractive hybridization is more expensive than RNase H-based methods. In most subtractive hybridization methods the capturing probes are biotinylated and the magnetic beads are streptavidin-coated, both of which are costly. For example, in the DOE tests and analyses reviewed earlier in the chapter, the streptavidin-coated magnetic beads alone accounted for 62% of the total reagent cost (David & Jensen, 2023). The cost difference becomes even more pronounced when larger sets of probes are used to enhance the probe coverage. In addition, compared with subtractive hybridization kits, RNase H-based kits allow a lower RNA input quantity, typically ranging from 10 to 1000 ng of total RNA. For instance the NEBNext kit has been successfully used to enrich indoor surface RNA viromes from total RNA inputs as low as 0.1 ng (Shiwa et al., 2023). Furthermore, in RNase H-based methods, rRNA depletion can occur either before or after cDNA synthesis. The depletion that occurs after cDNA synthesis allows for *in vitro* transcription using the cDNA as templates, leading to the amplification of total RNA after multiple rounds of transcription. An example is the so-called "post hoc RNase H digestion" method discussed later in the chapter (Wang et al., 2023) and listed in **Table 6.1**. The amplification enables RNase H-based depletion in applications with extremely low RNA inputs, such as single-cell RNA-seq (discussed later in this chapter). The described cases suggest RNase H-based methods are gaining popularity.

Clustered regularly interspersed short palindromic repeats/Cas9-based rRNA depletion

The CRISPR/Cas9-based (or short as Cas9-based) rRNA depletion method also relies on sequences complementary to rRNA. Clustered regularly interspersed short palindromic repeats (CRISPR) and CRISPR-associated (Cas) proteins are part of prokaryotic adaptive immune systems. Using this system bacteria and archaea cleave the Page 125 DNA of invading viruses and incorporate viral sequence fragments into the host's genome. By storing (or "remembering") the viral information bacteria and archaea can recognize past viral infections and use the Cas proteins to cleave specific locations on the viral genome if the same virus invades again. The CRISPR/Cas9 system has been engineered as a powerful genome editing tool, where Cas9, an endonuclease, cuts both strands of target DNA after its complexed single guide RNA (sgRNA) hybridizes to the target sequences. By designing specific sequences in the sgRNA this tool was adapted to selectively cleave unwanted high-abundance DNA sequences in the library, thereby preventing their downstream amplification and enriching for desired (uncut) sequences, and this approach was referred to as DASH (depletion of abundant sequences by hybridization) (Gu et al., 2016).

The CRISPR/Cas9-based DASH method was later adapted to remove microbial rRNA by using a pool of sgRNA designed for the bacteria of interest (Schmidt et al., 2019). This method was further optimized to achieve effective rRNA depletion with significantly lower RNA inputs (0.4–6 ng vs 500–1000 ng) and reduced costs ($3–7 vs $80 per sample) compared with the Ribo-Zero kit (Prezza et al., 2020). It includes a preamplification of cDNA with a few PCR cycles to boost overall RNA abundance. After the rRNA-derived cDNA is cleaved by Cas9, a postamplification with 12–24 cycles of PCR enriches mRNA-derived sequences. With the pre- and postamplification steps the minimal tested RNA input was as low as 0.4 ng.

Jumpcode Genomics, Inc. commercialized the CRISPR/Cas9-based depletion method. Their CRISPRclean Plus Stranded Total RNA Prep with rRNA Depletion (Human, Mouse, Rat, Pan-Bacteria) kit includes a pan-bacterial sgRNA pool that targets the 5, 16 and 23 S rRNA sequences of 212

bacterial species representing most bacterial phyla. This kit is suitable for low-input samples (5–100 ng of starting RNA). When applied to human nasopharyngeal or bronchoalveolar lavage samples, CRISPRclean increased pathogen genome coverage, especially for samples with low pathogen loads (Cerón et al., 2023). When compared to the RNase H-based Ribo-Zero Plus kit on 5 and 50 ng human fecal RNA inputs, CRISPRclean Plus demonstrated higher depletion efficiencies for both human and bacterial rRNA, leading to pathogen detection sensitivity and specificity comparable to the highly sensitive RT-qPCR method (Chan et al., 2023). Notably a higher fidelity was achieved with the 50 ng input (R^2 = 0.994) than with the 5 ng input (R^2 = 0.898) in this study. Page 126

Compared with methods/kits that remove rRNA before cDNA synthesis, CRISPR/Cas9 cuts unwanted sequences after cDNA synthesis or after cDNA amplification by various methods (**Table 6.1**), making it suitable for extremely low-input samples. This is advantageous in applications like single-cell RNA-seq discussed later in this chapter.

Blocking reverse transcription of rRNA in library preparation

An alternative approach employing sequences complementary to rRNA is to use them as blocking oligos to inhibit reverse transcription of rRNA. The QIAseq FastSelect 5S/16S/23S kit from QIAGEN employs highly modified oligonucleotides that bind to rRNA molecules and block the reverse transcriptase enzyme, thus preventing cDNA synthesis from rRNA. The design of the FastSelect oligonucleotides was based on the large numbers of 5, 16, and 23 S rRNA sequences in the SILVA database, and QIAGEN predicted>95% rRNA depletion of known rRNA sequences based on an in-silico modeling, making it suitable for complex communities from diverse environments. Further, the oligo design allows its compatibility with low-quality and highly fragmented RNA samples. Because rRNA depletion occurs during reverse transcription (following an optional fragmentation step) as an

integrated step in sequencing library preparation, it does not involve additional procedures for hybrid capturing or enzyme digestion and therefore it is fast (as short as 14 minutes), streamlined, and suitable for low-input samples (starting total RNA as low as 10 ng). In addition, a customizable option is available for sample-specific application. Owing to its effectiveness and simplicity, the QIAGEN kit is gaining popularity. In a comparative study on *Clostridium autoethanogenum*, the QIAGEN kit was evaluated alongside with a selective priming-based kit from NuGEN and a cDNA library normalization-based kit from Zymo Research (both discussed later in the chapter). The QIAGEN kit outperformed the other two kits in producing transcriptome profiles most consistent with previously published data, exhibiting more accurate strandedness, and detecting a wider range of transcript abundance distribution (de Lima et al., 2022).

Enrichment of mRNA by blocked rRNA

EMBR-seq (enrichment of mRNA by blocked rRNA) uses antisense rRNA sequences as blocking primers to prevent amplification of Page 127 rRNA-derived sequences by *in vitro* transcription (IVT), thereby achieving mRNA enrichment (Wangsanuwat et al., 2020). This method begins with annealing specific blocking primers at the 3′ end of the 5 , 16, and 23 S rRNA molecules. The RNA-primer hybrid at the 3′ end prevents subsequent polyadenylation of rRNA, as poly(A) polymerase preferentially adds adenines to single-stranded RNA, thereby allowing the selective polyadenylation of non-rRNA transcripts. After poly-A tails are added to non-rRNA, reverse transcription is conducted using a poly-T primer. The overhang of this poly-T primer incorporates a T7 promoter for IVT, a sample-specific barcode for multiplexing, and the 5′ Illumina adapter for sequencing library preparation. Following the synthesis of the second strand, only cDNA derived from non-rRNA is *in vitro* transcribed, while cDNA derived from rRNA lacking a T7 promoter, is not, enabling the amplification of non-rRNA through IVT. In experiments with *E. coli*, mRNA reads accounted for 90% of total mapped reads with EMBR-seq,

and transcript abundance was highly correlated (Pearson r = 0.97) with the no-depletion control (Wangsanuwat et al., 2020).

EMBR-seq is effective even with low-input RNA, with the lowest tested input being 0.02 ng, as mRNA enrichment is realized through IVT amplification after cDNA synthesis. In addition it requires only one to a few blocking primers per rRNA species (one at the 3′ end and a few optional at experimentally identified "hotspots"). Therefore this method substantially reduces the cost compared with many subtractive hybridization or RNase H-based methods that demand hundreds of probes for extensive probe tiling along rRNA molecules (Choe et al., 2021).

Ideally rRNA primers with a broad targeting range would enable the depletion of most rRNA in complex microbial communities. However designing such broad-range primers at the 3′ end can be challenging. Moreover, as this method relies primarily on a single blocking primer at the 3′ end of mature rRNA, it may be less effective for pre-rRNAs or degraded and fragmented rRNA that lacks the blocking primer annealing sites. Nevertheless a test on fragmented RNA indicated that its effectiveness was comparable to that of intact RNA (Wangsanuwat et al., 2020), though further investigation into its performance on degraded RNA is warranted. An enhanced version of this method, termed EMBR-seq+ was later developed by combining EMBR-seq with RNase H digestion for improved effectiveness (Heom et al., 2023) and is discussed in the "Combination of Methods" section. Page 128

Biased 3′ polyadenylation against rRNA

Amara and Vijaya (1997) observed selective polyadenylation of *E. coli* mRNA but not rRNA in intact polysomes using a yeast-derived poly(A) polymerase, suggesting that the 3′ ends of rRNAs are sterically blocked from polyadenylation when rRNAs are in complex with ribosomal proteins. Based on this observation Wendisch et al. (2001) developed a bacterial mRNA enrichment method using biased polyadenylation toward mRNA followed by purification with oligo(dT) chromatography. They tested this method on *E.*

coli and observed increased mRNA signals with transcript profiles comparable to those from total RNA without depletion. Besides oligo(dT) chromatography-enabled purification polyadenylated mRNAs can also be selectively converted to cDNA using oligo(dT) primers and then *in vitro* transcribed and amplified to further boost the mRNA enrichment effect; the same approach was applied by Frias-Lopez et al. (2008) to enrich mRNA from nanogram quantities of marine microbial community RNA. They reduced rRNA to 53% of the total sequences and maintained the relative abundance of most mRNAs, compared with unamplified transcriptome profiles. The authors suggested that the enrichment resulted not only from preferential polyadenylation but also likely from the high degree of rRNA secondary structure, which hindered their efficient amplification (Frias-Lopez et al., 2008). However, as Wilhelm and Landry (2009) pointed out, bias against 5′ ends and especially against long transcripts was expected when priming with oligo(dT) primer as compared to random oligomers in reverse transcription, since reverse transcriptase is well known to fall off templates as it progresses.

Selective priming in cDNA synthesis

Usually cDNA synthesis is primed with random oligomers (most frequently hexamers) for “unbiased” transcriptome coverage. Theoretically, using a subset of these entirely random oligomers that discriminates against rRNAs, one can selectively synthesize cDNA from mRNA. Based on the observation that many bacterial and archaeal rRNA sequences are richer in GC compared with mRNA and that TTTT sequences are found at a much higher frequencies in mRNA coding genes, Gonzalez and Robb (2007) applied HDTTTT and DHTTTT hexamers to prime cDNA synthesis from *Methanocaldococcus jannaschii* and *Pyrococcus horikoshii* and improved their Page 129 mRNA detection by PCR. This method is easy to conduct in the laboratory and does not require high RNA inputs.

Although the primer design was intended to be “universal” for bacteria and archaea, the differential occurrence of TTTT in non-rRNA versus in rRNA

varies broadly among microorganisms, causing variation in removal effectiveness and associated bias. Therefore more sophisticated selective primers with higher discriminating power were later designed computationally. By aligning all possible hexamers to rRNA and non-rRNA genes, Armour et al. (2009) selected a subset of hexamers without perfect match to rRNA while providing sufficient coverage to non-rRNA (referred to as "not-so-random primers"). This approach was applied to *Rhodopseudomonas palustris* and out of the total 4096 possible hexamers it identified 925 without perfect match to rRNA (Hirakawa et al., 2011). However, nonuniform coverage even within a transcript was observed, due to reduced randomness in priming of cDNA synthesis (Hirakawa et al., 2011). Therefore this method is not suitable for applications such as transcription start site detection and operon relationship determination, where coverage uniformity is required.

A commercial kit with generic selective primers, Ovation Prokaryotic RNA-seq System, was developed by NuGENTechnologies (later acquired by the Tecan Group in 2018) (Head et al., 2011). The not-so-random primers were designed against a collection of sequences from strains representing major bacterial and archaeal phyla and therefore predicted to be effective for a wide range of prokaryotes and even for partially degraded RNA. However, when tested on different bacterial monocultures, the effectiveness of this kit varied widely (Giannoukos et al., 2012; Peano et al., 2013). Another feature of this kit is that the double-stranded cDNA can be further amplified by PCR, which allows sequencing from very low quantities of input RNA. However, this may introduce well-known biases associated with exponential amplification, therefore undermining the accurate quantitation of mRNA abundance. For example, when applied to a *Burkholderia* sp., the correlation of mRNA abundance between transcriptomes with and without rRNA removal, as indicated by R^2, was only 0.51, much lower than the 0.96 obtained by MICROBExpress (Peano et al., 2013). In another evaluation Ovation Prokaryotic RNA-seq System was tested together with other four methods on

three well-characterized bacterial species with varied genome GC content. The subtractive hybridization kits achieved the highest fidelity (R^2 = 0.90–0.97 for MICROBExpress and R^2 = 0.88–0.95 for Ribo-Zero), followed by the Page 130 cDNA library normalization method discussed below (R^2 = 0.13–0.98). Ovation Prokaryotic RNA-seq System obtained much lower fidelity (R^2 = 0.43–0.53), only higher than the 5′ phosphate-dependent exonuclease digestion method, for which R^2 values were 0.04 to 0.42 (Giannoukos et al., 2012). A newer version of this kit, Universal Prokaryotic RNA-Seq Prokaryotic AnyDeplete by Tecan, also provides customizable option for species-specific rRNA removal.

cDNA library normalization

Ribosomal RNA can also be eliminated after cDNA synthesis through cDNA library normalization, a technique previously used to "equalize" transcripts to recover rare ones (Ko, 1990). Double-stranded cDNA is first denatured at an elevated temperature and then allowed to reanneal at a lower temperature. Because the reannealing rate of a cDNA is proportional to the square of its concentration, cDNAs from abundant transcripts (mostly rRNA and tRNA) return to double-stranded forms faster than those from less abundant transcripts. The reannealing is terminated after an appropriate time period so that most rRNA-derived cDNAs are double-stranded, while cDNAs from mRNAs remain single-stranded. The double-stranded cDNA can then be removed by enzymatic degradation with a duplex-specific nuclease (DSN) (Yi et al., 2011) or through physical separation with hydroxyapatite chromatography (VanderNoot et al., 2012). For example applying cDNA library normalization with DSN on *E. coli* reduced rRNA representation in the total sequences from 94% to 26%, while preserving mRNA relative abundance, with only the 10 most abundant mRNA transcripts exhibiting a small reduction after DSN treatment (Yi et al., 2011). As the depletion occurs after cDNA library construction, the RNA input can be lowered to

submicrogram quantities [e.g., 200–300 ng (VanderNoot et al., 2012; Yi et al., 2011)].

However the robustness of this method needs further investigation. As the reannealing rate is concentration-dependent, a fixed reannealing time might not work for samples with varied concentrations of rRNA and mRNA, and this might partly explain the observed variation of rRNA removal efficiency among samples (Yi et al., 2011). Furthermore the effectiveness is also impacted by genome GC content, as indicated by the high efficiencies and fidelity of rRNA removal from bacteria with low- and medium-GC content and the method's failure on high-GC bacterium (Giannoukos et al., 2012). This suggested that the Page 131 high-GC fraction of mRNA may anneal faster or form hairpins that are degraded by DSN, leading to the underrepresentation of mRNA transcripts (Giannoukos et al., 2012). Another adverse factor is the fact that within a transcriptome, mRNA abundance differs by several orders of magnitude. Although the equalization/normalization facilitates the detection of rare transcripts, it reduces the dynamic range of mRNA abundance and thus compromises quantitativeness. The effectiveness and fidelity may become even more difficult to achieve for environmental samples with community members differing in GC content and relative abundance. Conceivably rRNA from low-abundance organisms and highly expressed mRNA from abundant organisms may be at the same concentration range, making it difficult to differentiate rRNA and mRNA. Commercial kits based on this method are listed in **Table 6.1**.

Combination of methods

Due to the limitations of individual methods, a single method alone may not provide sufficient rRNA removal in some cases, particularly for challenging organisms or complex microbial communities. Therefore researchers have explored combining different methodologies to improve mRNA enrichment. For example, a coastal water metatranscriptome study employed 5′-

monophosphate-dependent exonuclease digestion followed by subtractive hybridization and reduced rRNA sequences to 37% of total reads (Poretsky et al., 2009). When systematically evaluated, the combination of these two methods provided more efficient rRNA removal than used alone, but produced much greater bias in mRNA abundance regardless of the order in which they were applied (He et al., 2010).

In contrast to the increased bias by combining subtractive hybridization and exonuclease digestion Peano et al. (2013) observed an interesting phenomenon when combining subtractive hybridization and selective priming. When removing rRNA from a high-GC *Burkholderia* sp. MICROBExpress alone, although offering a very high fidelity, was not efficient, while Ovation alone reduced rRNA to 61% of total reads but introduced a significant bias in mRNA. Interestingly applying MICROBExpress followed by Ovation not only lowered rRNA to 54% of total reads but also significantly enhanced fidelity as compared to using Ovation alone. This suggests that the application of MICROBExpress before Ovation made the latter less prone to bias or mitigated the bias introduced by the latter Page 132 (Peano et al., 2013). The underlying mechanism behind this bias reduction is unclear and needs further investigation for its broader application on high-GC microorganisms.

A more recent method, termed EMBR-seq+, combines the EMBR-seq approach with RNase H digestion to improve rRNA depletion efficiency (Heom et al., 2023). This method begins with annealing one to a few blocking primers to rRNA, which enables selective polyadenylation of non-rRNA and their subsequent preferential amplification through IVT. Then the IVT-amplified RNA (aRNA) is hybridized with a few 50 nt ssDNA probes designed to target strategic locations on rRNA for the subsequent degradation of rRNA-derived aRNA:DNA duplexes by RNase H.

EMBR-seq+ demonstrated up to 99% rRNA removal efficiencies when tested on several bacterial species and bacterial–fungal cocultures, showing high correlations to the no-depletion controls (Pearson r = 0.94–0.95). The

removal efficiencies surpassed those obtained by either EMBR-seq or RNase H methods used independently. Furthermore EMBR-seq+ requires less than 10 oligonucleotides (including both blocking primers and probes for RNase H) per rRNA species, making it cost-effective compared with methods that require extensive probe tiling along the rRNA molecules (Heom et al., 2023).

In general combining methods may improve rRNA depletion efficiency in certain cases. However the additional treatment and purification procedures can lead to greater material loss and increase the risk of mRNA degradation in some cases, given that mRNA is less stable than rRNA and tRNA. These considerations are important when designing combined approaches for rRNA depletion.

Ribosomal RNA depletion in dual RNA-seq

"Dual RNA-seq," which allows for simultaneous sequencing of the transcriptomes of both the host and associated microbiota, is a powerful approach to study host–microbial symbiotic or pathogenic interactions. In this application it is desirable to simultaneously deplete both host and microbial rRNA in one single step to streamline procedures and RNA handling, thereby reducing the risk of RNA degradation and lowering overall costs. Subtractive hybridization, RNase H-mediated depletion, rRNA reverse transcription blocking, and CRISPR/Cas9-based methods are suitable for this purpose, as probes, blocking oligos, and sgRNA targeting both hosts and microbes can be conveniently combined for codepletion in one single step. Page 133

Several commercial kits have been developed to target both eukaryotic and bacterial rRNAs. For instance the now-discontinued Illumina Ribo-Zero Gold rRNA Removal Kit (the Epidemiology version) utilized subtractive hybridization with probes targeting both Human/Mouse/Rat (HMR) and bacteria. This kit was widely applied, including its application to the marine sponge *Amphimedon queenslandica* holobiont (Xiang et al., 2022) and *Salmonella*-infected human cells (Westermann et al., 2016). Its replacement,

the RNase H-based Ribo-Zero Plus rRNA Depletion Kit also targets both HMR and bacteria and has been applied to study the human microbiome (Tan et al., 2023). Similarly, the QIAGEN QIAseq FastSelect Epidemiology Kit that blocks reversion transcription of rRNA from both HMR and bacteria was applied on mouse lung samples infected with *Pseudomonas aeruginosa* in a study of mouse infection models (Duncan et al., 2024). Additionally the CRISPRclean Plus Stranded Total RNA Prep with rRNA Depletion (Human, Mouse, Rat, Pan-Bacteria) kit offers an effective alternative for simultaneous removal of host and microbial rRNA, especially for samples with low RNA input, as demonstrated using human fecal RNA samples (Chan et al., 2023).

Alternatively researchers can combine probes that separately target hosts and microbes to achieve codepletion (Hayward et al., 2021; Shilpha et al., 2024; Marsh et al., 2017). Several manufacturers offer separate versions for eukaryotic and prokaryotic removal, but the different versions share very similar chemistry and protocols, making it convenient to combine probes. For exampleQIAseq FastSelect 5S/16S/23S kit was combined with QIAseq FastSelect Human Mouse Rat kit to study endogenous microbial symbionts of fish gills (François-Étienne et al., 2023). The subtractive hybridization-based siTOOLs kit allows users to combine up to four of their riboPOOLs. The combination of their Human and Pan-Prokaryote riboPOOLs at a 10:1 ratio has been used to investigate bacterial infections in lower extremity ulcers (Fritz et al., 2022). Similarly New England Biolabs recommends mixing probes from their NEBNext Human/Mouse/Rat kit with their Bacteria kit at ratio optimized to reflect the relative amounts of mammalian and bacterial rRNA in the sample. For example combining these two kits on a sample from filarial nematodes and *Wolbachia* endosymbionts lowered both host and bacterial rRNA reads and further depletion was achieved by using the NEBNext Custom RNA Depletion Design Tool to design probes targeting the nematode host (Cantin et al., 2024).

Ribosomal RNA depletion in single-cell RNA-seq

Traditional RNA-seq (or bulk RNA-seq) uses RNA extracted from the bulk population or community and therefore provides a snapshot of gene expression averaged from the entire population. In contrast, single-cell RNA-seq (scRNA-seq) investigates gene expression at the individual cell level and reveals cellular heterogeneity within a population. In scRNA-seq individual cells are placed into 96-well or 384-well plates using techniques such as fluorescence-activated cell sorting, or encapsuled in droplets using droplet microfluidics for higher throughput.

Despite advancements in scRNA-seq technology, similar to bulk RNA-seq, the overrepresentation of rRNA reads remains a challenge and scRNA-seq can also benefit from rRNA depletion. A single bacterial cell contains approximately 0.1 pg of RNA (Milo & Phillips, 2015), making it impractical to directly apply most rRNA removal methods developed for bulk RNA-seq to single cells. To address the overrepresentation of rRNA in scRNA-seq, two strategies have been developed: (1) *in situ* hybridization with specific probes for rRNA digestion or mRNA enrichment and (2) post hoc depletion following cDNA synthesis and amplification. Methods in these two categories are discussed below.

In situ RNase H digestion: In a droplet-based technology for bacterial scRNA-seq called BactDrop (Ma et al., 2023) RNase H-based rRNA depletion was performed *in situ* using the NEBNext rRNA Depletion Kit after cell fixation and permeabilization. This approach enables in-cell rRNA depletion before downstream processes. The fraction of mRNA reads among total aligned reads increased significantly, from ~5% in the no-depletion control to 50%–90% in the rRNA-depleted libraries, and mRNA profiles with and without rRNA depletion were highly correlated (Ma et al., 2023).

In situ hybridization with mRNA-specific probes for mRNA enrichment: Another droplet-enabled scRNA-seq method is called probe-based bacterial RNA-seq (ProBac-seq), and it employs a microfluidic device from the 10X

Genomics Single Cell 3′ system and large sets of genome-wide mRNA-specific ssDNA probes to selectively enrich for mRNA (McNulty et al., 2023; Samanta et al., 2024). In this method ssDNA probes are hybridized *in situ* to target mRNA within fixed and permeabilized cells, and individual cells are then encapsulated into droplets. As the probes also contain a 5′ PCR handle, a unique barcode, and a 3′ poly-A tail, barcoded cDNA can be generated and Page 135 amplified from the mRNA:probe hybrids via in-droplet PCR to achieve mRNA enrichment and amplification. Because sequencing libraries are constructed from the hybridized probes, only targeted mRNAs are sequenced. However this approach requires designing and generating large sets of oligonucleotide probes to target nearly the entire genome, therefore requiring prior genome information for probe design and entailing a significant upfront cost for probe synthesis.

Post hoc Cas9-based depletion after cDNA synthesis and amplification: Another approach is to amplify total RNA before rRNA depletion. For instance cDNA amplification using random displacement amplification sequencing (RamDA-seq) followed by Cas9-based depletion has been shown to significantly enhance detection sensitivity for bulk RNA samples with inputs as low as 0.2 pg or for single-sorted bacterial cells (Nishimura et al., 2023). Alternatively cDNA can also be amplified through PCR. In one example the Cas9 method was integrated into a scRNA-seq pipeline called MATQ-seq. In this approach bacterial cells are sorted into 96-well plates, lysed, and cDNA is synthesized using random primers and further amplified via PCR. The resulting rRNA-derived cDNA amplicon is then removed by using the Cas9 method (Homberger et al., 2023). In another example, a droplet-based scRNA-seq, called smRandom-seq, employs random primers for *in situ* cDNA generation and a microfluidic droplet barcoding platform for encapsulating individual cells into separate droplets, where a unique molecular index (or barcode) is added to each cell (Xu et al., 2023). Droplets are then disrupted to release barcoded cDNA for its amplification through PCR. Following amplification Cas9-based rRNA depletion is performed before generating the

sequencing libraries. These approaches enhanced mRNA detection sensitivity and allowed for a more efficient usage of sequencing data.

Post hoc RNase H digestion after *in vitro* transcription: As in BacDrop the RNase H method was also applied in another droplet-based single-cell transcriptome profiling called massively parallel, multiplexed, microbial sequencing (M3-seq) (Wang et al., 2023). But unlike BacDrop, where RNase H digestion occurs within the cell, M3-seq employs a post hoc RNase H digestion on the amplified RNA from *in vitro* transcription, aiming to minimize the unwanted loss of unamplified mRNA transcripts, as the authors previously noted that *in situ* rRNA depletion by RNase H can reduce mRNA capture efficiency (Wang et al., 2023). In this method, after two rounds of indexing, double-stranded cDNA synthesis, and PCR amplification, DNA Page 136 libraries are *in vitro* transcribed to single-stranded RNA. The rRNA-derived sequences in the amplified libraries are then hybridized to rRNA-specific DNA probes and digested with RNase H for removal. Compared with libraries without rRNA depletion, this method achieved 11- to 27-fold increase in mRNA reads with minimal bias in mRNA profile (Pearson r = 0.94), and the authors also demonstrated that this approach resulted in a significantly higher rRNA depletion efficiency than the post hoc Cas9-based approach (Wang et al., 2023).

Conclusions and perspectives

Significant progress has been made in depleting rRNA from total RNA in microbial transcriptomic analysis. Each depletion method comes with its own strengths and limitations. The choice of method depends on various factors, such as sample characteristics (e.g., GC content, genome complexity, phylogenic composition, and RNA quantity and quality) as well as application-specific characteristics, which may differ in the requirements for input RNA quality and quantity, quantitativeness in mRNA abundance, uniformity in transcript coverage, representation of rare transcripts, etc.

Methods that utilize sequences complementary to rRNA are the most prevalent in the field. Among them, subtractive hybridization was the first to emerge and has been the most widely used approach, owing to its high removal efficiency and low bias when using properly designed probes, along with various commercial kits. Meanwhile the more recently developed methods within this category, such as RNase H- and CRISPR/Cas9-based depletion, are gaining popularity for their effectiveness with low-input RNA quantities and their adaptability for single-cell RNA-seq. In these antisene rRNA sequence-based methods customized probes (or sgRNA) can be employed when samples are not adequately targeted by generic commercial probes, owing to various tools developed for designing probes against user-supplied sequences and custom options provided with commercial kits. Moreover a range of DIY methods has been developed to provide customizable, cost-effective alternatives to commercial kits. Methods such as biased polyadenylation, selective priming, and cDNA library normalization are likely less constrained by sample phylogenetic composition and RNA quantity and quality. However these methods should be employed with caution, as potential biases associated with them are not yet fully understood. Page 137

For environmental samples with complex community compositions, such as soil microbiota, using chemically synthesized short sequences to tile rRNA for effective depletion would require prohibitively large pools of probes. In such cases, using microbial community gDNA as templates to generate sample- or community-specific long probes could serve as a viable option for subtractive hybridization or RNase H-mediated digestion. Given that the rRNA content is generally lower in these environments compared with fast-growing laboratory cultures, sequencing the entire transcriptome without rRNA depletion, known as “total RNA-seq,” may be a practical alternative. This approach has been applied to simultaneously assess the taxonomic composition of active microbiota and their expressed functions in various environments, including soil (Bang-Andreasen et al., 2020; Urich et al., 2008;

Tveit et al., 2014), cow rumen (Li & Guan, 2017), and rhizosphere (Turner et al., 2013). As sequencing throughput continues to increase and costs decrease, alongside the development of bioinformatics tools that effectively utilize rRNA read data, rRNA depletion may become less critical for RNA-seq analysis of complex community samples in the future.

References

Amara and Vijaya, 1997 Amara RR, Vijaya S. Specific polyadenylation and purification of total messenger RNA from *Escherichia coli*. *Nucleic Acids Research*. 1997;25(17):3465–3470 **https://doi.org/10.1093/nar/25.17.3465**.

Armour et al., 2009 Armour CD, Castle JC, Chen R, et al. Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nature Methods*. 2009;6(9):647–649 **https://doi.org/10.1038/nmeth.1360**.

Azarani and Hecker, 2000 Azarani A, Hecker KH. RNA chromatography under thermally denaturing conditions: Analysis and quality determination of RNA. *ADS Biotec Application Note*. 2000;116:1–4.

Bang-Andreasen et al., 2020 Bang-Andreasen T, Anwar MZ, Lanzén A, et al. Total RNA sequencing reveals multilevel microbial community changes and functional responses to wood ash application in agricultural and forest soil. *FEMS Microbiology Ecology*. 2020;96 **https://doi.org/10.1093/femsec/fiaa016**.

Bhagwat et al., 2014 Bhagwat A, Ying Z, Smith A. Evaluation of ribosomal RNA removal protocols for *Salmonella* RNA-Seq projects. *Advances in Microbiology*. 2014;4:25–32 **http://dx.doi.org/10.4236/aim.2014.41006**.

Cantin et al., 2024 Cantin LJ, Gregory V, Blum LN, Foster JM. Dual RNA-seq in filarial nematodes and Wolbachia endosymbionts

using RNase H based ribosomal RNA depletion. *Frontiers in Microbiology*. 2024;15 **https://doi.org/10.3389/fmicb.2024.1418032**.

Castro et al., 2013 Castro TLP, Seyffert N, Ramos RTJ, et al. Ion torrent-based transcriptional assessment of a *Corynebacterium pseudotuberculosis* equi strain reveals denaturing high-performance liquid chromatography a promising rRNA depletion method. *Microbial Biotechnology*. 2013;6(2):168–177 **https://doi.org/10.1111/1751-7915.12020**.

Cerón et al., 2023 Cerón S, Clemons NC, Von Bredow B, Yang S. Application of CRISPR-based human and bacterial ribosomal RNA depletion for SARS-CoV-2 shotgun metagenomic sequencing. *American Journal of Clinical Pathology*. 2023;159(2):111–115 **https://doi.org/10.1093/ajcp/aqac135**.

Chan et al., 2023 Chan AP, Siddique A, Desplat Y, et al.... A CRISPR-enhanced metagenomic NGS test to improve pandemic preparedness. *Cell Reports Methods*. 2023;3 **https://doi.org/10.1016/j.crmeth.2023.100463**.

Choe et al., 2021 Choe D, Szubin R, Poudel S, et al. RiboRid: A low cost, advanced, and ultra-efficient method to remove ribosomal RNA for bacterial transcriptomics. *PLoS Genetics*. 2021;17(9):e1009821 **https://doi.org/10.1371/journal.pgen.1009821**.

Ciulla et al., 2010 Ciulla D, Giannoukos G, Earl A, et al. Evaluation of bacterial ribosomal RNA (rRNA) depletion methods for sequencing microbial community transcriptomes. *Genome Biology*. 2010;11(Suppl 1):P9 **https://doi.org/10.1186/1465-6906-11-S1-P9**.

Culviner et al., 2020 Culviner PH, Guegler CK, Laub MT, Cooper VS. A simple, cost-effective, and robust method for rRNA depletion in

RNA-sequencing studies. *mBio*. 2020;11 **https://doi.org/10.1128/mbio.00010-20**.

David and Jensen, 2023 David BM, Jensen PA. Improving an rRNA depletion protocol with statistical design of experiments. *SLAS Technology*. 2023;28(1):16–21 **https://doi.org/10.1016/j.slast.2022.09.004**.

de Lima et al., 2022 de Lima LA, Reinmets K, Nielsen LK, et al. RNA-seq Sample Preparation Kits Strongly Affect Transcriptome Profiles of a Gas-Fermenting Bacterium. *Microbiology Spectrum*. 2022;10(4):e02303–e02322 **https://doi.org/10.1128/spectrum.02303-22**.

Duncan et al., 2024 Duncan RP, Moustafa DA, Lewin GR, et al. Improvement of a mouse infection model to capture Pseudomonas aeruginosa chronic physiology in cystic fibrosis. *Proceedings of the National Academy of Sciences*. 2024;121e2406234121 **https://doi.org/10.1073/pnas.2406234121**.

Dunman et al., 2001 Dunman PM, Murphy E, Haney S, et al. Transcription profiling-based identification of *Staphylococcus aureus* genes regulated by the agr and/or sarA Loci. *Journal of Bacteriology*. 2001;183(24):7341–7353 **https://doi.org/10.1128/jb.183.24.7341-7353.2001**.

Engelhardt et al., 2020 Engelhardt F, Tomasch J, Häussler S. Organism-specific depletion of highly abundant RNA species from bacterial total RNA. *Access Microbiology*. 2020;2:acmi000159 **https://doi.org/10.1099/acmi.0.000159**.

Filiatrault, 2011 Filiatrault MJ. Progress in prokaryotic transcriptomics. *Current Opinion in Microbiology*. 2011;14(5):579–586 **https://doi.org/10.1016/j.mib.2011.07.023**.

François-Étienne et al., 2023 François-Étienne S, Nicolas L, Eric N, et al. Important role of endogenous microbial symbionts of fish gills in the challenging but highly biodiverse Amazonian blackwaters. *Nature Communications*. 2023;143903 **https://doi.org/10.1038/s41467-023-39461-x**.

Frias-Lopez et al., 2008 Frias-Lopez J, Shi Y, Tyson GW, et al. *Microbial community gene expression in ocean surface waters*. *Proceedings of the National Academy of Sciences of the United States of America*. 2008;105(10):3805–3810 **https://doi.org/10.1073/pnas.0708897105** United States.

Fritz et al., 2022 Fritz BG, Kirkegaard JB, Nielsen CH, Kirketerp-Møller K, Malone M, Bjarnsholt T. Transcriptomic fingerprint of bacterial infection in lower extremity ulcers. *APMIS: Acta Pathologica, Microbiologica, et Immunologica Scandinavica*. 2022;130(8):524–534 **https://doi.org/10.1111/apm.13234**.

Giannoukos et al., 2012 Giannoukos G, Ciulla DM, Huang K, et al. Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biology*. 2012;13 **https://doi.org/10.1186/gb-2012-13-3-r23**.

Gonzalez and Robb, 2007 Gonzalez JM, Robb FT. Counterselection of prokaryotic ribosomal RNA during reverse transcription using non-random hexameric oligonucleotides. *Journal of Microbiological Methods*. 2007;71(3):288–291 **https://doi.org/10.1016/j.mimet.2007.09.010**.

Gu et al., 2016 Gu W, Crawford ED, O'Donovan BD, et al. Depletion of abundant sequences by hybridization (DASH): Using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biology*. 2016;17 **https://doi.org/10.1186/s13059-016-0904-5**.

Hausen and Stein, 1970 Hausen P, Stein H. Ribonuclease H an enzyme degrading the RNA moiety of DNA-RNA hybrids.

European Journal of Biochemistry. 1970;14(2):278–283 **https://doi.org/10.1111/j.1432-1033.1970.tb00287.x**.

Hayward et al., 2021 Hayward RJ, Humphrys MS, Huston WM, Myers GSA. Dual RNA-seq analysis of in vitro infection multiplicity and RNA depletion methods in *Chlamydia*-infected epithelial cells. *Scientific Reports*. 2021;11 **https://doi.org/10.1038/s41598-021-89921-x**.

He et al., 2010 He S, Wurtzel O, Singh K, et al. Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nature Methods*. 2010;7(10):807–812 **https://doi.org/10.1038/nmeth.1507**.

Head et al., 2011 Head SR, Komori HK, Hart GT, et al. Method for improved illumina sequencing library preparation using NuGEN ovation RNA-Seq system. *Biotechniques*. 2011;50(3):177–181 **https://doi.org/10.2144/000113613**.

Heom et al., 2023 Heom KA, Wangsanuwat C, Butkovich LV, et al. Targeted rRNA depletion enables efficient mRNA sequencing in diverse bacterial species and complex co-cultures.. *mSystems*. 2023;8(6):e0028123 **https://doi.org/10.1128/msystems.00281-23**.

Hirakawa et al., 2011 Hirakawa H, Oda Y, Phattarasukol S, et al. *Activity of the Rhodopseudomonas palustris p-coumaroyl-homoserine lactone-responsive transcription factor RpaR*. *Journal of Bacteriology*. 2011;193(10):2598–2607 **https://doi.org/10.1128/JB.01479-10** United States.

Homberger et al., 2023 Homberger C, Hayward RJ, Barquist L, Vogel J. *Improved bacterial single-cell RNA-Seq through automated MATQ-Seq and Cas9-based removal of rRNA reads*. *mBio*. 2023;14 e03557-22 **https://doi.org/10.1128/mbio.03557-22**.

Huang et al., 2020 Huang Y, Sheth RU, Kaufman A, Wang HH. Scalable and cost-effective ribonuclease-based rRNA depletion for transcriptomics. *Nucleic Acids Research*. 2020;48(4):E20 **https://doi.org/10.1093/nar/gkz1169**.

Jiang and Belasco, 2004 Jiang X, Belasco JG. Catalytic activation of multimeric RNase E and RNase G by 5′-monophosphorylated RNA. *Proceedings of the National Academy of Sciences of the United States of America*. 2004;101(25):9211–9216 **https://doi.org/10.1073/pnas.0401382101**.

Karpinets et al., 2006 Karpinets TV, Greenwood DJ, Sams CE, Ammons JT. *RNA: Protein ratio of the unicellular organism as a characteristic of phosphorous and nitrogen stoichiometry and of the cellular requirement of ribosomes for protein synthesis*. *BMC Biology*. 2006;4 **https://doi.org/10.1186/1741-7007-4-30** United States.

Kim et al., 2019 Kim IV, Ross EJ, Dietrich S, Döring K, Alvarado AS, Kuhn CD. Efficient depletion of ribosomal RNA for RNA sequencing in planarians. *BMC Genomics*. 2019;20 **https://doi.org/10.1186/s12864-019-6292-y**.

Ko, 1990 Ko MSH. An 'equalized cDNA library' by the reassociation of short double-stranded cDNAs. *Nucleic Acids Research*. 1990;18(19):5705–5711 **https://doi.org/10.1093/nar/18.19.5705**.

Kukutla et al., 2013 Kukutla P, Steritz M, Xu J. *Depletion of ribosomal RNA for mosquito gut metagenomic RNA-seq*. *Journal of Visualized Experiments* 2013; e50093 **https://doi.org/10.3791/50093**.

Li and Guan, 2017 Li F, Guan LL. Metatranscriptomic profiling reveals linkages between the active rumen microbiome and feed efficiency in beef cattle. *Applied and Environmental*

Microbiology. 2017;83 **https://doi.org/10.1128/AEM.00061-17**.

Li et al., 2013 Li SK, Zhou JW, Yim AKY, et al. *Organism-specific rRNA capture system for application in next-generation sequencing*. *PLoS One*. 2013;8 **https://doi.org/10.1371/journal.pone.0074286** Hong Kong.

Lucas et al., 2023 Lucas JK, Gruenke PR, Burke DH. *Minimizing amplification bias during reverse transcription for in vitro selections*. *RNA*. 2023;29(8):1301–1315 **https://doi.org/10.1261/rna.079650.123** Epub 2023 May 16. PMID: 37192815; PMCID: PMC10351896.

Ma et al., 2023 Ma P, Amemiya HM, He LL, et al. Bacterial droplet-based single-cell RNA-seq reveals antibiotic-associated heterogeneous cellular states. *Cell*. 2023;186(4):877 **https://doi.org/10.1016/j.cell.2023.01.002**.

Marsh et al., 2017 Marsh JW, Humphrys MS, Myers GSA. A laboratory methodology for dual RNA-sequencing of bacteria and their host cells in vitro. *Frontiers in Microbiology*. 2017;8:01830 **https://doi.org/10.3389/fmicb.2017.01830**.

McGrath et al., 2008 McGrath KC, Thomas-Hall SR, Cheng CT, et al. Isolation and analysis of mRNA from environmental microbial communities. *Journal of Microbiological Methods*. 2008;75(2):172–176 **https://doi.org/10.1016/j.mimet.2008.05.019**.

McNulty et al., 2023 McNulty R, Sritharan D, Pahng SH, et al. Probe-based bacterial single-cell RNA sequencing predicts toxin regulation. *Nature Microbiology*. 2023;8(5):934–945 **https://doi.org/10.1038/s41564-023-01348-4**.

Milo and Phillips, 2015 Milo R, Phillips R. *Cell biology by the numbers* Garland Science 2015;

https://doi.org/10.1201/9780429258770.

Moran, 2009 Moran MA. Metatranscriptomics: Eavesdropping on complex microbial communities. *Microbe*. 2009;4:329–335 **http://dx.doi.org/10.1128/microbe.4.329.1**.

Neidhardt and Umbarger, 1996 Neidhardt FC, Umbarger HE. *Chemical composition of Escherichia coli, 2nd ed, Escherichia coli and Salmonella: Cellular and molecular biology* Washington, DC: ASM Press; 1996;.

Nishimura et al., 2023 Nishimura M, Takeyama H, Hosokawa M. Enhancing the sensitivity of bacterial single-cell RNA sequencing using RamDA-seq and Cas9-based rRNA depletion. *Journal of Bioscience and Bioengineering*. 2023;136:152–158 **https://doi.org/10.1016/j.jbiosc.2023.05.010**.

Pang et al., 2004 Pang X, Zhou D, Song Y, et al. Bacterial mRNA purification by magnetic capture-hybridization method. *Microbiology and Immunology*. 2004;48(2):91–96 **https://doi.org/10.1111/j.1348-0421.2004.tb03493.x**.

Pastor et al., 2022 Pastor MM, Sakrikar S, Rodriguez DN, Schmid AK. Comparative analysis of rRNA removal methods for RNA-Seq differential expression in halophilic archaea. *Biomolecules*. 2022;12 **https://doi.org/10.3390/biom12050682**.

Peano et al., 2013 Peano C, Pietrelli A, Consolandi C, et al. An efficient rRNA removal method for RNA sequencing in GC-rich bacteria. *Microbial Informatics and Experimentation*. 2013;3 **https://doi.org/10.1186/2042-5783-3-1**.

Petrov et al., 2022 Petrov DP, Kaiser S, Kaiser S, Jung K. Opportunities and challenges to profile mRNA modifications in *Escherichia coli*. *Chembiochem: A European Journal of Chemical Biology*. 2022;23 **https://doi.org/10.1002/cbic.202200270**.

Petrova et al., 2017 Petrova OE, Garcia-Alcalde F, Zampaloni C, Sauer K. Comparative evaluation of rRNA depletion procedures for the

improved analysis of bacterial biofilm and mixed pathogen culture transcriptomes. *Scientific Reports*. 2017;7 **https://doi.org/10.1038/srep41114**.

Plum and Clark-Curtiss, 1994 Plum G, Clark-Curtiss JE. Induction of Mycobacterium avium gene expression following phagocytosis by human macrophages. *Infection and Immunity*. 1994;62(2):476–483 **https://doi.org/10.1128/iai.62.2.476-483.1994**.

Poretsky et al., 2009 Poretsky RS, Hewson I, Sun S, Allen AE, Zehr JP, Moran MA. Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environmental Microbiology*. 2009;11(6):1358–1375 **https://doi.org/10.1111/j.1462-2920.2008.01863.x**.

Prezza et al., 2020 Prezza G, Heckel T, Dietrich S, Homberger C, Westermann AJ, Vogel J. Improved bacterial RNA-seq by Cas9-based depletion of ribosomal RNA reads. *RNA (New York, N.Y.)*. 2020;26(8):1069–1078 **https://doi.org/10.1261/rna.075945.120**.

Rauhut and Klug, 1999 Rauhut R, Klug G. mRNA degradation in bacteria. *FEMS Microbiology Reviews*. 1999;23(3):353–370 **https://doi.org/10.1111/j.1574-6976.1999.tb00404.x**.

Robinson et al., 1994 Robinson KA, Robb FT, Schreier HJ. Isolation of maltose-regulated genes from the hyperthermophilic archaeum, *Pyrococcus furiosus*, by subtractive hybridization. *Gene*. 1994;148:137–141 **https://doi.org/10.1016/0378-1119(94)90247-X**.

Samanta et al., 2024 Samanta P, Cooke SF, McNulty R, Hormoz S, Rosenthal A. ProBac-seq, a bacterial single-cell RNA sequencing methodology using droplet microfluidics and large oligonucleotide probe sets. *Nature Protocols*. 2024;19(10):2939–2966 **https://doi.org/10.1038/s41596-024-01002-1**.

Schmidt et al., 2019 Schmidt ST, Yu FB, Blainey PC, May AP, Quake SR. Nucleic acid cleavage with a hyperthermophilic Cas9 from an uncultured Ignavibacterium. *Proceedings of the National Academy of Sciences of the United States of America*. 2019;116(46):23100–23105 **https://doi.org/10.1073/pnas.1904273116**.

Shilpha et al., 2024 Shilpha J, Lee J, Kwon JS, et al. An improved bacterial mRNA enrichment strategy in dual RNA sequencing to unveil the dynamics of plant-bacterial interactions. *Plant Methods*. 2024;20 **https://doi.org/10.1186/s13007-024-01227-x**.

Shiwa et al., 2023 Shiwa Y, Baba T, Sierra MA, Kim JK, Mason CE, Suzuki H. Evaluation of rRNA depletion methods for capturing the RNA virome from environmental surfaces. *BMC Research Notes*. 2023;16 **https://doi.org/10.1186/s13104-023-06417-9**.

Shrestha et al., 2009 Shrestha PM, Kube M, Reinhardt R, Liesack W. Transcriptional activity of paddy soil bacterial communities. *Environmental Microbiology*. 2009;11(4):960–970 **https://doi.org/10.1111/j.1462-2920.2008.01821.x**.

Sooknanan et al., 2010 Sooknanan R, Pease J, Doyle K. *Novel methods for rRNA removal and directional, ligation-free RNA-seq library preparation*. *Nature Methods*. 2010;7:i–ii (2010) **https://doi-org/10.1038/nmeth.f.313**.

Sorek and Cossart, 2010 Sorek R, Cossart P. Prokaryotic transcriptomics: A new view on regulation, physiology and pathogenicity. *Nature Reviews Genetics*. 2010;11(1):9–16 **https://doi.org/10.1038/nrg2695**.

Stewart et al., 2010 Stewart FJ, Ottesen EA, DeLong EF. Development and quantitative analyses of a universal rRNA-subtraction

protocol for microbial metatranscriptomics. *The ISME Journal*. 2010;4:896–907 **https://doi.org/10.1038/ismej.2010.18**.

Su and Sordillo, 1998 Su C, Sordillo LM. A simple method to enrich mRNA from total prokaryotic RNA. *Molecular Biotechnology*. 1998;10(1):83–85 **https://doi.org/10.1007/bf02745865**.

Tan et al., 2023 Tan A, Murugapiran S, Mikalauskas A, et al. Rational probe design for efficient rRNA depletion and improved metatranscriptomic analysis of human microbiomes. *BMC Microbiology*. 2023;23 **https://doi.org/10.1186/s12866-023-03037-y**.

Turner et al., 2013 Turner TR, Ramakrishnan K, Walshaw J, et al. Comparative metatranscriptomics reveals kingdom level changes in the rhizosphere microbiome of plants. *ISME Journal*. 2013;7(12):2248–2258 **https://doi.org/10.1038/ismej.2013.119**.

Tveit et al., 2014 Tveit AT, Urich T, Svenning MM. Metatranscriptomic analysis of arctic peat soil microbiota. *Applied and Environmental Microbiology*. 2014;80:5761–5772 **https://doi.org/10.1128/AEM.01030-14**.

Urich et al., 2008 Urich T, Lanzén A, Qi J, Huson DH, Schleper C, Schuster SC. *Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. PLoS One*. 2008;3 **https://doi.org/10.1371/journal.pone.0002527** Norway.

VanderNoot et al., 2012 VanderNoot VA, Langevin SA, Solberg OD, et al. *cDNA normalization by hydroxyapatite chromatography to enrich transcriptome diversity in RNA-seq applications*. *Biotechniques*. 2012;53(6):373–380 **https://doi.org/10.2144/000113937** United States.

Wahl et al., 2022 Wahl A, Huptas C, Neuhaus K. *Comparison of rRNA depletion methods for efficient bacterial mRNA sequencing*.

Scientific Reports. 2022;12 article number 5765 **https://doi.org/10.1038/s41598-022-09710-y**.

Wang et al., 2023 Wang B, Lin AE, Yuan J, et al. Single-cell massively-parallel multiplexed microbial sequencing (M3-seq) identifies rare bacterial populations and profiles phage infection. *Nature Microbiology*. 2023;8(10):1846–1862 **https://doi.org/10.1038/s41564-023-01462-3**.

Wangsanuwat et al., 2020 Wangsanuwat C, Heom KA, Liu E, O'Malley MA, Dey SS. Efficient and cost-effective bacterial mRNA sequencing from low input samples through ribosomal RNA depletion. *BMC Genomics*. 2020;21 **https://doi.org/10.1186/s12864-020-07134-4**.

Wendisch et al., 2001 Wendisch VF, Zimmer DP, Khodursky A, Peter B, Cozzarelli N, Kustu S. Isolation of *Escherichia coli* mRNA and comparison of expression using mRNA and total RNA on DNA microarrays. *Analytical Biochemistry*. 2001;290(2):205–213 **https://doi.org/10.1006/abio.2000.4982**.

Westermann et al., 2016 Westermann AJ, Förstner KU, Amman F, et al. Dual RNA-seq unveils noncoding RNA functions in host-pathogen interactions. *Nature*. 2016;529(7587):496–501 **https://doi.org/10.1038/nature16547**.

Wilhelm and Landry, 2009 Wilhelm BT, Landry JR. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods (San Diego, Calif.)*. 2009;48(3):249–257 **https://doi.org/10.1016/j.ymeth.2009.03.016**.

Wurtzel et al., 2010 Wurtzel O, Sapra R, Chen F, Zhu Y, Simmons BA, Sorek R. *A single-base resolution map of an archaeal transcriptome. Genome Research*. 2010;20(1):133–141 **https://doi.org/10.1101/gr.100396.109** Israel.

Xiang et al., 2022 Xiang X, Poli D, Degnan BM, Degnan SM. Ribosomal RNA-depletion provides an efficient method for successful dual RNA-Seq expression profiling of a marine sponge holobiont. *Marine Biotechnology*. 2022;24(4):722–732 **https://doi.org/10.1007/s10126-022-10138-8**.

Xu et al., 2023 Xu Z, Wang Y, Sheng K, et al.... Droplet-based high-throughput single microbe RNA sequencing by smRandom-seq. *Nature Communications*. 2023;14 **https://doi.org/10.1038/s41467-023-40137-9**.

Yi et al., 2011 Yi H, Cho YJ, Won S, et al. Duplex-specific nuclease efficiently removes rRNA for prokaryotic RNA-seq. *Nucleic Acids Research*. 2011;39 **https://doi.org/10.1093/nar/gkr617**.

Zhao et al., 1999 Zhao J, Hyman L, Moore C. Formation of mRNA 3' ends in eukaryotes: Mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiology and Molecular Biology Reviews*. 1999;63:405–445 **https://doi.org/10.1128/mmbr.63.2.405-445.1999**.

CHAPTER 7

Anything, anyone, anywhere: metagenomic sequencing with Oxford Nanopore Technologies

Jessica L. Davies, Julia Eales, Caz Harrold and Aaron Pomerantz, Oxford Nanopore Technologies plc, Gosling Building, Edmund Halley Road, Oxford Science Park, Oxford, United Kingdom

Abstract

The study of complex microbial communities through metagenomics has been greatly enhanced by advances in DNA sequencing technologies. This chapter explores the significant impact of Oxford Nanopore Technologies on the field of metagenomics and how researchers are utilizing the unique capabilities of the technique to gain detailed insights into microbial genomes. After an introduction to how nanopore sequencing works and its benefits for metagenomic sequencing the chapter walks through the sample-to-answer metagenomic sequencing workflow, from sample preparation to data analysis. Finally case studies highlight the use of nanopore metagenomic sequencing in applications ranging from infectious disease research to the analysis of air and soil microbiomes.

Keywords

Nanopore; Oxford Nanopore; sequencing; clinical research; infectious disease; environmental research; microbial sequencing; microbiome

Introduction

Over the past two decades advances in DNA sequencing technologies have revolutionized the field of metagenomics, allowing scientists to analyze the genetic material of entire microbial communities directly from samples. These advances have made it possible to obtain complete or nearly complete genome sequences from microbes that cannot be cultured in the laboratory, providing invaluable insights into their biology, ecology, and evolution (Dickson, 2021; Shi et al., 2022).

Oxford Nanopore has significantly impacted the field of metagenomics with its unique sequencing capabilities (Agustinho et al., 2024). The vision of Oxford Nanopore Technologies is to enable the analysis of anything, by anyone, anywhere. Researchers around the world are leveraging metagenomic nanopore sequencing to push the boundaries of microbial research, across wide-ranging applications spanning from clinical research and pathogen surveillance to uncovering the vast diversity of our planet's microbial life.

How does Oxford Nanopore sequencing work?

At the heart of Oxford Nanopore sequencing technology lies the nanopore itself: a tiny pore that functions as a gateway, facilitating the passage of molecules from one side of a membrane to the other (**Fig. 7.1**). In Oxford Nanopore sequencing protein nanopores are used; these are small protein channels. In nature nanopores vary in their structure and properties; these features can be further enhanced through precise mutagenesis to improve sequencing performance. Page 144 Oxford Nanopore has selected those with optimal features for nanopore sequencing.

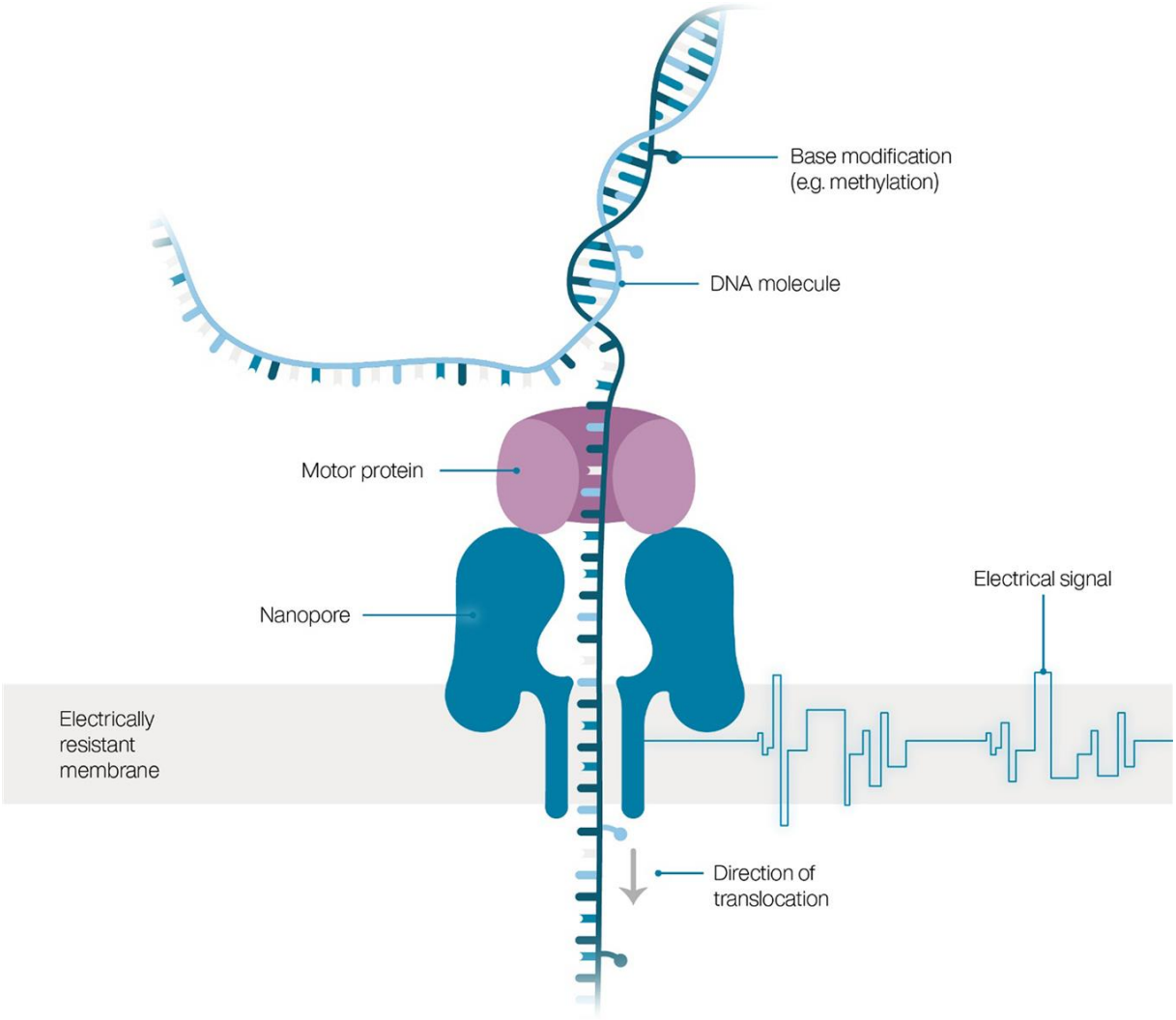


FIGURE 7.1 How Oxford Nanopore sequencing works. In Oxford Nanopore sequencing a strand of DNA (or RNA) passes through a nanopore embedded in an electrically resistant membrane. Its movement is controlled by the motor protein, which is bound to adapters that are attached to the DNA or RNA molecule before sequencing. This motor protein "unzips" the double-stranded molecule and feeds a single strand through the nanopore. The passage of the nucleic acid strand through the nanopore causes a measurable disruption in the ionic current. This is measured by an electrode as an electrical signal and used to determine the sequence of bases.

The diagram shows a D N A molecule being threaded through a nanopore embedded in an electrically resistant membrane. A motor protein is illustrated controlling the movement of the D N A through the nanopore. As the D N A translocates through the nanopore, an electrical signal is generated, which is represented by a graph with peaks and troughs to the right of the nanopore. The diagram also indicates a base modification (example: methylation) on the D N A molecule, which can be detected by changes in the electrical signal as the D N A passes through the nanopore.

In nanopore sequencing nanopores are embedded in an electro-resistant membrane, within individual wells in an array. Each nanopore corresponds to its own electrode, connected to a sensor chip. When an electrical potential is applied to this system, ions can flow through the nanopore, producing a measurable current. All these elements are enclosed within a flow cell—the consumable component of the nanopore sequencing process within which the sequencing itself takes place (**Fig. 7.2**). The flow cell is placed inside a nanopore sequencing device, and a prepared DNA or RNA sample is added to the flow cell before starting a sequencing run. A sequencing run is started using Page 145 software onboard the sequencer itself, or via a laptop or computer to which the device is plugged in.



FIGURE 7.2 An Oxford Nanopore MinION™ Flow Cell. The flow cell is the consumable component of the nanopore sequencing process. The sensor array is made up of sensing wells, each collecting data from a nanopore. The DNA or RNA sample to be sequenced is added to the sensor array via the sample port; any excess fluid then accumulates in the waste channel. For more information on nanopore sequencing devices see section **Nanopore sequencing**.

The device shown is rectangular in shape and black in color. The labeled parts of the device are as follows: A 'waste channel' which appears to be a groove or conduit for the disposal of waste material. A 'sample port cover' which seems to be a protective covering for the area where a sample would be introduced into the device and it contains text labeled as Spot ON. A 'sensor array' which is present below the sample port cover. The device also contains a serial number labeled as F A Q 79154 and a Q R code. Screws are visible in the device at various points.

To prepare a sample of native DNA or RNA for nanopore sequencing an adapter—to which an enzyme known as a motor protein is bound—is added to the end of each molecule (**Fig. 7.1**). Oxford Nanopore sequencing is the only commercially available technology that enables the direct sequencing of native RNA molecules, without the need to convert the RNA to complementary DNA (cDNA; also known as copy DNA). When preparing native RNA libraries for direct RNA sequencing first-strand synthesis is recommended before this step, producing a double-stranded molecule comprised of one RNA and one cDNA strand; however only the native RNA strand is sequenced.

When a prepared molecule of DNA or RNA approaches a nanopore, the motor protein "unzips" the double-stranded molecule and controls the movement of one strand through the hollow core of the nanopore. As this strand passes through, it causes a distinct disruption in the flow of ions; this is measured as an electrical signal by the electrode associated with that nanopore. This signal is then converted into the sequence of DNA or RNA bases using a basecalling algorithm. Epigenetic modifications in native DNA and RNA, such as methylation, also produce characteristic electrical signals and therefore can also Page 146 be detected (**Fig. 7.1**). When the strand has completely passed through the nanopore, that nanopore is then free to begin sequencing another strand.

This process is highly scalable, with sequencing proceeding in parallel across hundreds or thousands of nanopores simultaneously. Several types of flow cells are available; each option contains a different number of nanopores, providing a range of outputs to suit the needs of different sequencing experiments. Each type of flow cell pairs with a range of nanopore sequencing devices, providing two levels of scalability: the output of an individual flow cell plus the number of flow cells that can be run in parallel on a single device. Within the range of sequencing devices are portable sequencers that can be plugged into a laptop, for sequencing anywhere (see section **On-demand metagenomic sequencing—in the laboratory or**

in the field) and ultra-high-throughput sequencers that can generate terabases of sequencing data.

How does nanopore sequencing enhance metagenomic research?

Complete metagenome-assembled genomes and comprehensive variant detection from a single technology

Short-read sequencing technologies require DNA and RNA samples to be fragmented during library preparation, producing sequencing reads that typically span only a few hundred base pairs (bp). This process can often lead to incomplete genome assemblies, particularly when dealing with metagenomic samples. Repetitive DNA sequences and closely related microorganisms are especially challenging to resolve accurately due to the limited length of reads, making it difficult to reconstruct complex genomic regions.

In contrast Oxford Nanopore sequencing does not require fragmentation—nanopores can process any length of DNA or RNA fragment presented to them, producing reads of any length, from short to ultra-long (Oxford Nanopore Technologies, 2024a). The current longest recorded read exceeds 4 million bases [4 megabases (Mb)] in length (Oxford Nanopore Technologies, 2024b). This capability enhances the reconstruction of complete genome sequences for each microorganism present in a metagenomic sample, known as metagenome-assembled genomes (MAGs), including by spanning repetitive regions and structural variants in the genome—leading to more accurate and contiguous microbial genome assemblies, including for

Page 147

plasmids (**Fig. 7.3**). The genomes of some microbes can even be sequenced end to end in single reads, removing the need for assembly entirely (Beaulaurier et al., 2020).

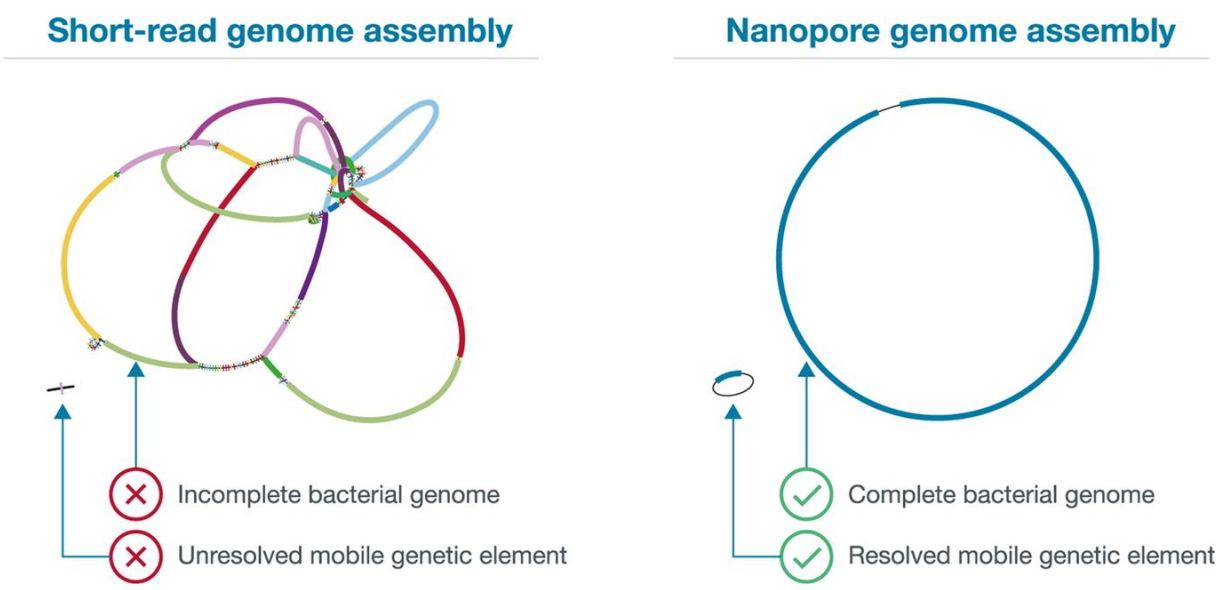


FIGURE 7.3 Visual comparison of bacterial genome assembly between short-read and Oxford Nanopore sequencing technologies. Legacy short-read sequencing technologies generate short reads of<300 bp in length. As a result bacterial genomes are reconstructed using computational techniques; however, due to the short length of the reads, there are often regions of the genome that are not successfully sequenced—positional genomic information can be lost and mobile genetic elements left unresolved. The resulting genome assemblies are often incomplete and noncontiguous (left). In contrast nanopore reads, which can be of any length from short (20 bases) to ultra-long (>4 Mb) (Oxford Nanopore Technologies, 2024b), span larger regions of bacterial genomes in single reads, meaning that more of the genome is successfully sequenced and there is less reliance on computational reconstruction. The resulting bacterial genome assemblies are complete and contiguous, and mobile genetic elements are fully resolved (right).

On the left, the short-read genome assembly shows a complex, tangled structure in various colors with labels indicating incomplete bacterial genome

and unresolved mobile genetic element which is shown as a black line. On the right, the nanopore genome assembly shows a simple, complete circular structure in blue with a small black segment and labels indicating complete bacterial genome and resolved mobile genetic element which is a small circular ring with blue segment.

Nanopore reads enable the precise differentiation of even closely related species (Portik et al., 2022) and the accurate identification of microbial strains within mixed samples (Buytaers et al., 2024; Nicholls et al., 2019). The high accuracy of nanopore sequencing data ensures high-quality whole-genome assembly from nanopore reads only—without the need for assembly "polishing" using data from a second sequencing technology (Bogaerts et al., 2024; Hall et al., 2024; Lerminiaux et al., 2024; Liu et al., 2022; Sanderson et al., 2024; Sereika et al., 2022; Zhao et al., 2023; Oxford Nanopore Technologies, 2024c).

As well as resulting in complete MAGs nanopore reads can distinguish bacterial genome sequences from associated plasmids and produce complete, separate plasmid sequences. This is important in the detection of genes encoding antimicrobial resistance (AMR) and their location in either the bacterial genome or a plasmid (Sauerborn et al., 2024). Information on the location of AMR genes is critical, as those located in plasmids can be rapidly transferred via the process of horizontal gene transfer. Page 148

Nanopore sequencing does not require the use of PCR—a process that results in biased coverage of genomes and limits sequencing to regions that can be amplified, leaving gaps in genome assemblies. Instead nanopore sequencing processes native DNA and RNA strands directly, avoiding PCR bias and allowing microbial genomes to be sequenced in full, including GC-rich and highly repetitive regions that are not amenable to amplification (Aird et al., 2011; Pugh, 2023).

The use of PCR also means that epigenetic modifications, such as DNA and RNA methylation, are lost from nucleic acid samples, meaning that they must

instead be indirectly detected using special library preparation steps that can be time consuming and give inconsistent results. With PCR-free nanopore sequencing the distinct disruptions in current caused by epigenetic modifications enable their direct detection from native DNA and RNA, without the need for additional library preparation steps, providing more in-depth characterization of samples. Oxford Nanopore basecalling algorithms support the direct detection of a growing range of modifications in DNA and RNA; it is also possible for researchers to train their own basecalling algorithms to detect their modifications of interest. Using Oxford Nanopore sequencing researchers can study DNA methylation (Heidelbach et al., 2024; Wilbanks et al., 2022) an important epigenetic modification that affects gene expression and other cellular processes. In bacteria DNA methylation can protect against invading viruses (Casadesús & Low, 2006). Strain-specific methylation patterns can also be used to enhance resolution of closely related microorganisms from metagenomic sequencing data (Wilbanks et al., 2022).

Epigenetic data can also be used to support metagenomic contig binning—the process of grouping sequences from the same organism from within a mixed microbial sample—as well as to identify misassembled contigs and associate mobile genetic elements with their host genomes (Heidelbach et al., 2024).

Nanopore reads facilitate the detection not only of single nucleotide polymorphisms but also structural variants (SVs)—defined as spanning 50 bp or more—across microbial genomes, including in challenging regions such as long repetitive sequences. This comprehensive variant detection is important in the identification of strain-level differences between microbes, the characterization of known and novel pathogens, and understanding the molecular basis of environmental adaptation and disease mechanisms. Page 149

Real-time sequencing and rapid workflows

A key feature of Oxford Nanopore sequencing technology is the ability to sequence and analyze data in real time (Benner et al., 2007). Where legacy sequencing technologies feature a set run time, after which data become available for analysis, nanopore sequencing data can be basecalled and analyzed as sequencing progresses, allowing an experiment to be monitored and the sequencing run to be stopped as soon as enough data have been generated.

This real-time data streaming capability allows immediate access to sequencing results, which is crucial for applications requiring rapid responses, such as outbreak investigations, pathogen surveillance, and AMR analysis (Sauerborn et al., 2024). In addition a sequencing run can be quality checked as it proceeds and stopped once sufficient data have been generated to answer a biological question. Provided that sufficient active nanopores remain to meet the desired output, the flow cell can then be washed and reused for another sample.

Oxford Nanopore also provides rapid, streamlined, automatable library preparation options—with kits requiring as little as 10 minutes to take samples from extracted nucleic acids to sequencing-ready libraries—reducing turnaround times further still.

On-demand metagenomic sequencing—in the laboratory or in the field

The range of Oxford Nanopore sequencing devices offers scalability to suit the needs of the sequencing experiment—from portable devices through to ultra-high-throughput benchtop devices (Oxford Nanopore Technologies, 2024d).

Legacy sequencing technologies can require considerable infrastructure with a large footprint, confining sequencing to a laboratory environment. In addition samples often need to be batched into a single sequencing run or sent to a centralized facility for sequencing, in each case leading to potential waiting times of weeks or months. In contrast compact, portable Oxford

Nanopore devices enable metagenomic sequencing on site and in house or in the field—even in remote locations (**Fig. 7.4**) (Urban et al., 2023).



FIGURE 7.4 Portable Oxford Nanopore MinION sequencer. Portable Oxford Nanopore MinION sequencers enable sequencing anywhere—in the laboratory or in the field. Image credit: Aaron Pomerantz.

On-site metagenomic nanopore sequencing has been utilized in extreme environments, including for studies assessing the impact of climate change through off-grid analysis of soil samples from Europe's largest ice cap (Vatnajökull, Iceland) (Gowers et al., 2019) and Page 150 permafrost thaw analysis in Alaska (Haan & Drown, 2021; Seitz et al., 2021).

Bringing the sequencer to the sample source is essential for metagenomic sequencing applications requiring time-critical answers and ensures that samples do not need to be transported long distances, during which they may

degrade. With a flexible range of sequencers available, there is no need to wait to batch samples—they can instead be sequenced on demand, avoiding delays. This flexibility, combined with low startup costs, makes metagenomic nanopore sequencing an accessible and cost-effective option for researchers.

From sample to answer: the metagenomic Oxford Nanopore sequencing workflow

Sample extraction and library preparation

The metagenomic sequencing workflow begins with the extraction of high-quality DNA or RNA from samples. Oxford Nanopore provides internally validated extraction protocols for a wide range of sample types, including environmental water, stool, and soil. As nanopores sequence any-length fragments of native DNA and RNA present in a sample, selecting a method that preserves high molecular-weight nucleic acids will maximize outputs of long reads. Fragmentation and Page 151 size selection are not required but may be used to optimize for a specific read length range.

Next samples are prepared for nanopore sequencing by the addition of adapters during library preparation. Oxford Nanopore sequencing kits feature ligation-based methods of adapter addition, for a fully PCR-free workflow. Alternatively, if starting from low sample inputs, PCR-based methods may be used. Sample barcoding (indexing) kits are also available to prepare samples for sequencing together on one flow cell (known as “multiplexing”)—for example, where lower depth of coverage per sample is required (**Fig. 7.5**).

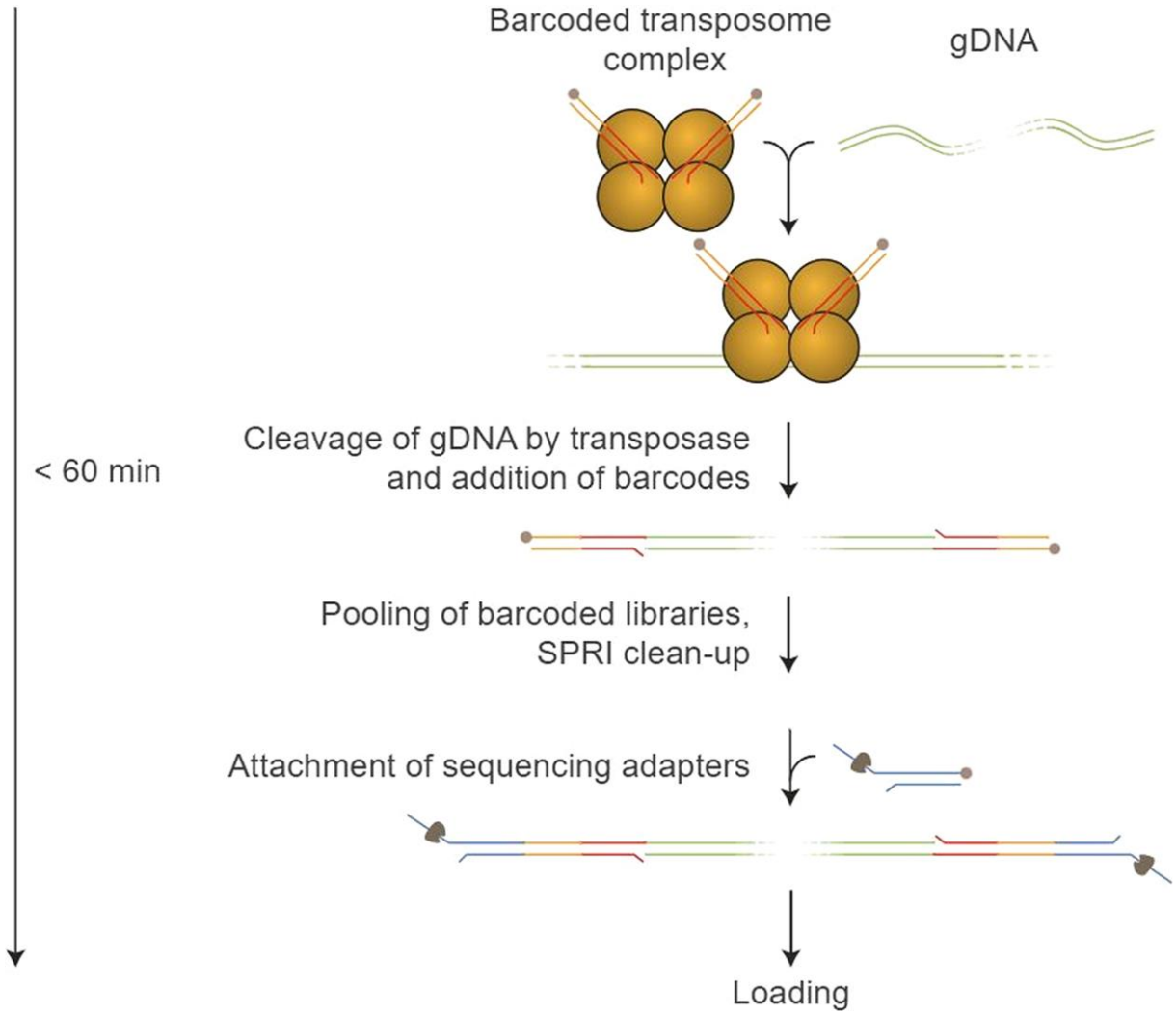


FIGURE 7.5 Example DNA library preparation workflow, here using the Rapid Barcoding Kit. In this workflow a barcoded transposome complex—comprising a transposase enzyme and barcoded tags—is used to simultaneously cleave genomic DNA (gDNA) molecules and attach barcodes to the cleaved ends. Each sample to be sequenced in multiplex is prepared using a different barcode; these are then pooled together to form a single sample. Finally the sequencing adapters, to which the motor protein is prebound, are attached to the DNA molecule ends via rapid attachment chemistry. The sample is now ready for sequencing on an Oxford Nanopore

device. This library preparation step takes around 1 hour and does not require PCR.

The process starts with the barcoded transposome complex and g D N A. The g D N A is cleaved by transposase and barcodes are added. Next, the barcoded libraries are pooled and cleaned up using S P R I. Sequencing adapters are then attached and the final step is loading. These steps take less than 60 minutes.

Where turnaround time is critical, rapid chemistry-based sequencing kits—for which PCR-free and PCR-based options are available—enable library preparation in as little as 10 minutes. In this technique a Page 152 transposase enzyme is used to cut the DNA and attach adapters in a single rapid step. Rapid sample barcoding kits are also available, in which this technique is used to add barcodes to samples before the addition of sequencing adapters (**Fig. 7.5**). This chemistry can also be used to linearize circular DNA inputs, such as plasmids. Alternatively circular DNA can be linearized enzymatically before library preparation.

Nanopore sequencing

The pocket-sized MinION device is ideal for metagenomic sequencing in any location. Scaling up for higher throughput needs the flexible, benchtop GridION™ sequencer provides sequencing on up to five flow cells at once. Every flow cell is individually addressable—sequencing runs can be separately started and stopped as needed on each flow cell without interrupting ongoing experiments, allowing multiple researchers to use the device as needed. Finally the PromethION™ devices allow sequencing on high-output PromethION Flow Cells, making them ideal for higher-depth sequencing of metagenomic samples, for example, to characterize low-abundance strains within a sample. A range of PromethION sequencers is available, each

providing a different flow cell capacity, to scale up to ultra-high-throughput sequencing requirements.

With real-time data streaming during sequencing, basecalling and data analysis can begin as soon as a run starts.

Metagenomic data analysis

A key goal of metagenomic sequencing is to identify the diverse microbes present within a sample through taxonomic classification. This typically involves comparing the DNA sequences from the sample to a reference database of known microbial DNA. Bioinformatics tools match each read from the sample against this database to find the best match, identifying the known microorganisms present. Identifying AMR genes in the same sample is important for tracking and combating antibiotic resistance. Bioinformatics tools achieve this by comparing sequencing data to databases of known AMR genes to detect and quantify resistance.

Another goal is the reconstruction of complete microbial genome sequences from metagenomic samples. This is achieved using de novo assembly: a process in which genome sequences are pieced together from the raw sequencing data without the need for a reference genome. This is especially useful for studying diverse microbial communities where existing reference genomes might not be available. Page 153 Bioinformatics tools align and merge nanopore reads to create near-complete MAGs. Once these genomes are assembled, they can be analyzed to identify genes, predict their functions, and understand the roles of different microbes within microbiomes. This comprehensive approach allows researchers to explore the genetic diversity and functional potential of microbial communities in detail —even in the absence of well-characterized reference genomes. The ability to generate long and ultra-long reads with Oxford Nanopore sequencing is significantly improving the reconstruction of high-quality MAGs from metagenomic samples today (Cuscó et al., 2022; Sereika et al., 2022; Simon et al., 2023).

Nanopore sequencing can also directly detect various types of DNA methylation commonly found in microbes, such as 5-methylcytosine (5mC), N6-methyladenine (6 mA), and N4-methylcytosine (4mC). This capability identifies specific methylation patterns in metagenomic samples, which aids metagenomic binning, detection of contamination in MAGs, and linking of mobile genetic elements, such as plasmids, to their microbial hosts (Beaulaurier et al., 2018; Cao et al., 2024). New data analysis tools are continuously being developed to use these methylation patterns to better understand microbial communities and their epigenetic modifications (Heidelbach et al., 2024).

To make bioinformatics tools more accessible to researchers Oxford Nanopore offers EPI2ME™, a user-friendly platform designed to streamline the analysis of nanopore sequencing data. The EPI2ME platform includes a metagenomics pipeline that identifies and quantifies microorganisms and detects potential AMR genes by comparing sequences to reference databases. The results are presented in various visual formats, such as abundance tables, interactive Sankey and sunburst plots showing relationships between microorganisms (**Fig. 7.6**), and bar plots comparing the most common microbes across samples. These visualizations help researchers to better understand the microbial diversity, abundance, and resistance patterns within their samples.

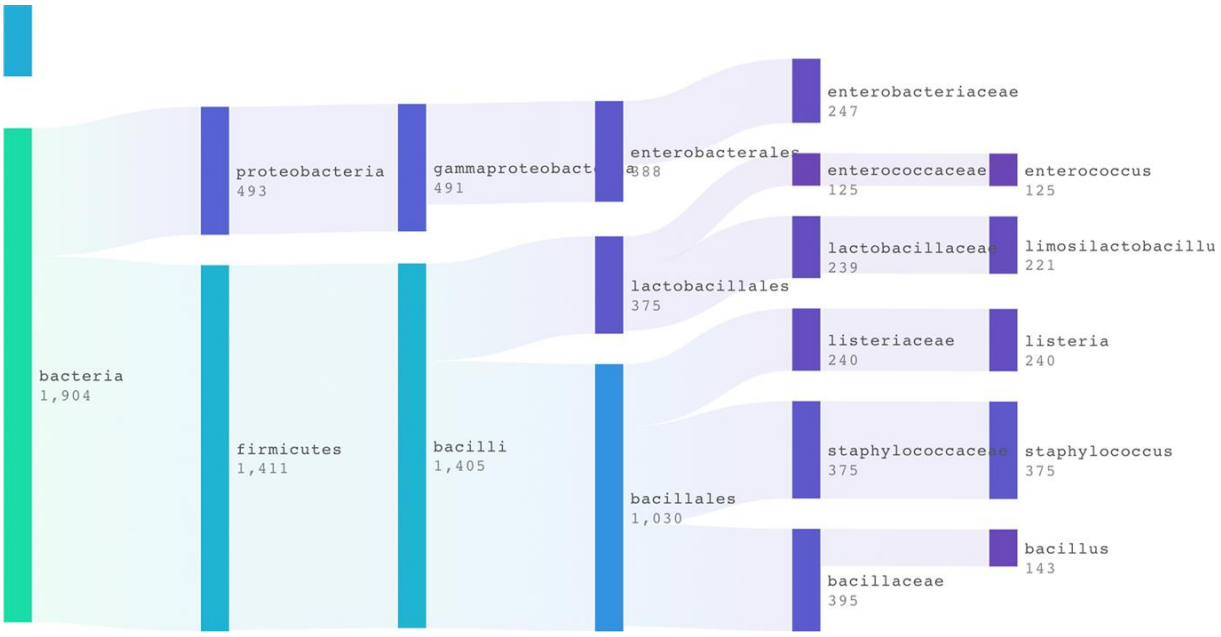


FIGURE 7.6 An example Sankey plot output from the Oxford Nanopore data analysis platform, EPI2ME. An example Sankey plot output from an EPI2ME bioinformatics pipeline showing bacterial taxonomical data from a metagenomic nanopore sequencing dataset. This workflow is accessible via both an intuitive graphical user interface and the command line and provides visualizations for taxonomy, diversity, and abundance from mixed microbial samples.

The flowchart starts with bacteria, which has a count of 1,904. It branches into two categories: proteobacteria with a count of 493 and firmicutes with a count of 1,411. Proteobacteria further branches into gammaproteobacteria with a count of 491, which then branches into enterobacterales with a count of 388. Enterobacterales further branches into enterobacteriaceae with a count of 247. Firmicutes branches into bacilli with a count of 1,405, which further branches into lactobacillales with a count of 375 and bacillales with a count of 1,030. Lactobacillales branches into enterococcaceae with a count of 125 and lactobacillaceae with a count of 239. Enterococcaceae leads to enterococcus with a count of 125. Lactobacillaceae branches into limosilactobacillu with a count of 221. Bacillales branches into listeriaceae

with a count of 240 and staphylococcaceae with a count of 375. Listeriaceae leads to listeria with a count of 240 and staphylococcaceae leads to staphylococcus with a count of 375. Bacillales also branches into bacillaceae with a count of 395, which leads to bacillus with a count of 143.

Metagenomic sequencing or 16S sequencing?

Although, by definition, metagenomics involves whole-genome analysis, targeted sequencing of the 16S ribosomal RNA (rRNA) gene—a widely used method of microbial identification since the advent of high-throughput next-generation sequencing—should also be considered in this context (Durazzi et al., 2021; Kleikamp et al., 2023). Page 154

The 16S rRNA gene is ~1.5 kb in length. In 16S sequencing using short-read technology read lengths are limited and only partial fragments of the variable regions of this gene can be sequenced, limiting the possible taxonomic resolution. In contrast nanopore sequencing can span the full 16S rRNA gene in single reads, providing accurate, species-level microbial identification (Zhang et al., 2023). Oxford Nanopore provides a streamlined, end-to-end 16S sequencing workflow, from library preparation through to sequencing and data analysis.

The end-to-end 16S nanopore sequencing workflow offers high taxonomic resolution for targeted bacterial and archaeal identification from mixed microbial communities. For experiments requiring comprehensive characterization of microbial genomes identification of rare pathogens, up to strain-level taxonomic resolution, and identification of AMR genes, variants, and epigenetic modifications, whole-genome metagenomic nanopore sequencing is recommended.

Applications of metagenomic nanopore sequencing for human and environmental health

The transformative potential of nanopore sequencing in respiratory metagenomics

Lower respiratory infections (LRIs) pose a significant health challenge worldwide, causing substantial mortality, morbidity, and economic Page 155 burden. According to the World Health Organization, LRIs were responsible for approximately 2.6 million deaths globally in 2019 (World Health Organization, 2024).

Current methods to diagnose LRIs, such as cell culture, are often slow and have poor sensitivity, impeding targeted therapy, and often leading to the overuse of broad-spectrum antibiotics, ultimately exacerbating the development and spread of AMR. The development of rapid and accurate diagnostic methods is therefore crucial to improve patient outcomes, improve antibiotic stewardship, and tackle this global threat (Charalampous et al., 2019). To this end Oxford Nanopore sequencing offers several advantages over alternative sequencing platforms for pathogen identification and AMR profiling, including rapid library preparation and real-time data acquisition.

Charalampous et al. (2019) published a pioneering study in 2019, demonstrating the potential of Oxford Nanopore technology in the rapid identification of bacterial LRIs. The team developed a nanopore sequencing workflow capable of identifying pathogens within 6 hours from sample collection to result. A critical step in their methodology was the application of a saponin-based technique for host DNA depletion, which removed up to 99.999% of human host nucleic acid. With the high proportion of host DNA present in respiratory samples this depletion step was crucial to be able to identify pathogens without a targeted, and therefore biased, approach, such as multiplex PCR.

The researchers applied their workflow to 40 research samples and then optimized and tested it on an additional 41 research samples, achieving highly accurate results. The authors suggested that this rapid and accurate method of bacterial LRI identification showed the potential to reduce the overuse of

broad-spectrum antibiotics and enable more tailored treatments for bacterial LRIs (Charalampous et al., 2019).

Building on this successful research Charalampous et al. (2023) performed a proof-of-concept study exploring the feasibility of implementing their rapid nanopore sequencing workflow in a routine clinical setting. Here the team evaluated the performance of a pilot respiratory metagenomics (RMg) service for patients in an intensive care unit with respiratory infections, alongside routine testing.

In total 128 samples were processed. Sequencing reports generated for 30-minute and 2-hour timepoints were assessed for pathogen identification, with AMR information included in the 2-hour reports. The median turnaround time from sample receipt to reporting was Page 156 6.7 hours, with 90% of samples receiving final reports on the same day. Furthermore, through the identification of between-patient transmission, their results impacted infection control. This study provides real-world evidence of the significant impact that a rapid nanopore-based metagenomic workflow could have on patient management and infection control (Charalampous et al., 2023).

The RMg workflow is limited to the analysis of DNA-containing microbes, as the saponin-based method is less effective at preserving RNA viruses. RNA viruses, such as influenza, SARS-CoV-2, and enteroviruses, are also important targets for respiratory infection analysis. Aware of this limitation the team have published an optimized workflow for simultaneous detection of DNA and RNA-containing microbes (Alcolea-Medina et al., 2024). This is achieved through a rapid, mechanical host depletion method, after which the RNA is converted to double-stranded DNA for subsequent sequencing. The authors obtained high accuracy for both bacterial and viral detection, after only 2 hours of nanopore sequencing. In prospective validation of the workflow on 33 research samples taken from patients with suspected pneumonia they detected additional pathogens in 21% of samples compared with what had been identified through routine tests (Alcolea-Medina et al., 2024).

These collaborative research efforts together demonstrate the significant potential that nanopore sequencing could have in future real-world clinical applications, through timely and accurate LRI pathogen identification, supporting targeted and appropriate antibiotic use, and enhanced infection control measures.

Exploring the depths of diversity in environmental microbiomes

Metagenomic sequencing can reveal complex relationships between human and environmental health, whether it is identifying potential pathogens in water, soil, or air samples, or monitoring the interaction between anthropogenic pressures, climate change, and the environmental microbiome (Urban et al., 2023). Nanopore technology has been used across continents to explore ecosystem microbiomes through sequencing a range of environmental samples, including those obtained from air and soil.

Airborne microorganisms constitute a crucial yet underexplored component of environmental microbiomes. Air sampling has gained attention for its potential to monitor public health threats and environmental changes. For example research published by Reska et al. (2023) Page 157 showcased the first study to use metagenomic nanopore sequencing to characterize the full diversity of the air microbiome. The study highlighted the limitations of previous approaches, which were often focused on targeted detection methods, and commonly involved amplification techniques that could introduce biases, leading to an incomplete picture of microbial diversity. In contrast amplification-free metagenomic nanopore sequencing provided a more comprehensive view, allowing for robust species identification even in ultra-low biomass samples like air (Reska et al., 2023).

Having first established that their nanopore sequencing workflow was suitable, and laboratory and computational approaches were optimal, 30 air samples were collected more than three weeks from five urban locations in Barcelona, Spain. The results demonstrated that microbial compositions were

highly location-specific, with distinct taxa characterizing different environments. For example the “Green Belt” area was distinguished by unique microbial taxa and the absence of the more dominant taxa found in urbanized areas. Despite the variation in microbial composition the study found no significant correlation between environmental variables, such as air pollution and measurements of anthropogenic impact, and microbial diversity. One notable finding was the detection of AMR genes across all sampled environments, with the *VanR-O g*ene, responsible for vancomycin resistance, being one of the most frequently detected (Reska et al., 2023).

Minor et al. (2023) similarly investigated the potential of nanopore sequencing of the air microbiome to improve public health surveillance, compared with the limited view obtained from pathogen-specific surveillance strategies. The team performed metagenomic nanopore sequencing of air samples to detect human respiratory and enteric viruses in various congregate settings. This study employed a method known as sequence-independent single-primer amplification to provide a broader view of the virome present in the air samples. They detected 13 different human RNA viruses, including those associated with respiratory and enteric diseases, in 19 of the 22 air samples collected. The authors suggested that a metagenomic analysis of air samples could be a powerful tool for real-time monitoring of viral pathogens (Minor et al., 2023).

Moving away from the air microbiome to that of the earth, soil hosts one of the most diverse microbial habitats on Earth. The health and productivity of plants, as well as the overall stability of ecosystems, are intimately linked to the microbial life in soil. Understanding these Page 158 complex interactions is particularly important in the face of environmental changes, such as those resulting from climate change, which has seen an increase in permafrost thaw and soil disturbance events; these phenomena can lead to shifts in microbial communities, producing significant ecological consequences (Seitz et al., 2021).

Haan & Drown (2021) have used nanopore sequencing to investigate the impact of permafrost thaw and soil disturbance on microbial and plant communities in soil samples from Alaska, United States (Seitz et al., 2021). By analyzing soil cores from different permafrost thaw gradients the researchers found that microbial communities from across the thaw gradient differentially impacted plant productivity, in terms of height and leaf number (Seitz et al., 2021). This was linked to a shift in the taxonomic composition and abundance of soil microbes, with the general finding that microbes associated with more disturbed soil negatively affected plant productivity. The study underscores the importance of understanding soil microbial dynamics, particularly in regions undergoing rapid environmental change, as these changes can have wide-reaching consequences on plant communities and ecosystem health (Seitz et al., 2021).

The team have also explored the relationship between soil disturbance and AMR in Alaskan soils (Haan & Drown, 2021). This research revealed that soil disturbance not only shifts microbial community composition but also increases the prevalence of AMR genes. The researchers found that isolates from disturbed soils exhibited higher resistance to clinically significant antibiotics. The resistance genes identified were predominantly associated with antibiotic efflux mechanisms; they also identified several resistance genes on plasmids, which could be easily disseminated. Their findings suggest that environmental changes, such as permafrost thaw, could facilitate the spread of AMR, posing a potential threat to both environmental and human health (Haan & Drown, 2021).

In a different yet related context Gowers et al. (2019) demonstrated the feasibility of conducting metagenomic sequencing in remote and challenging environments using nanopore technology. The team conducted an expedition to Europe's largest ice cap (Vatnajökull, Iceland), where they sequenced soil samples from a geothermal hot spring gorge using an entirely off-grid, solar-powered setup. The results revealed that the microbial community was dominated by cyanobacteria characteristic of polar regions and thermophilic

bacteria, indicative of the unique microenvironment (Gowers et al., 2019). This study Page 159 highlights the adaptability of nanopore sequencing in its ability to be applied in low-resource environments with minimal infrastructure, providing the possibility to explore microbial diversity in some of the most isolated and extreme environments on Earth.

The application of nanopore sequencing in environmental metagenomics has provided unparalleled insights into the microbial communities of air and soil. The research outlined here illustrates the power of Oxford Nanopore technology to unravel complex environmental interactions that have significant implications for both human and ecological health.

Conclusion

In summary Oxford Nanopore sequencing technology offers significant advantages for metagenomics, including enhanced whole-genome assembly from mixed microbial samples, direct detection of epigenetic modifications, real-time data access, portability, and cost-effectiveness. These benefits are transforming the ability to study and understand microbial communities, enabling rapid and detailed genomic analysis—both in the laboratory and in the field.

References

Agustinho et al., 2024 Agustinho DP, Fu Y, Menon VK, Metcalf GA, Treangen TJ, Sedlazeck FJ. Unveiling microbial diversity: Harnessing long-read sequencing technology. *Nature Methods*. 2024;21:954–966 **https://doi.org/10.1038/s41592-024-02262-1**.

Aird et al., 2011 Aird D, Ross MG, Chen W-S, et al. Analyzing and minimizing PCR amplification bias in illumina sequencing libraries. *Genome Biology*. 2011;12:R18 **https://doi.org/10.1186/gb-2011-12-2-r18**.

Alcolea-Medina et al., 2024 Alcolea-Medina A, Alder C, Snell LB, et al. Unified metagenomic method for rapid detection of microorganisms in clinical samples. *Communications Medicine*. 2024;4 **https://doi.org/10.1038/s43856-024-00554-3**.

Beaulaurier et al., 2020 Beaulaurier J, Luo E, Eppley JM, et al. Assembly-free single-molecule sequencing recovers complete virus genomes from natural microbial communities. *Genome Research*. 2020;30:437–446 **https://doi.org/10.1101/gr.251686.119**.

Beaulaurier et al., 2018 Beaulaurier J, Zhu S, Deikus G, et al. Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nature Biotechnology*. 2018;36:61–69 **https://doi.org/10.1038/nbt.4037**.

Benner et al., 2007 Benner S, Chen RJA, Wilson NA, et al. Sequence-specific detection of individual DNA polymerase complexes in real time using a nanopore. *Nature Nanotechnology*. 2007;2:718–724 **https://doi.org/10.1038/nnano.2007.344**.

Bogaerts et al., 2024 Bogaerts B, Van den Bossche A, Verhaegen B, et al. Closing the gap: Oxford Nanopore Technologies R10 sequencing allows comparable results to Illumina sequencing for SNP-based outbreak investigation of bacterial pathogens. *Journal of Clinical Microbiology*. 2024;62:e01576–23 **https://doi.org/10.1128/jcm.01576-23**.

Buytaers et al., 2024 Buytaers FE, Verhaegen B, Nieuwenhuysen TV, et al. Strain-level characterization of foodborne pathogens without culture enrichment for outbreak investigation using shotgun metagenomics facilitated with nanopore adaptive sampling. *Frontiers in Microbiology*. 2024;15 **https://doi.org/10.3389/fmicb.2024.1330814**.

Cao et al., 2024 Cao L, Kong Y, Fan Y, et al. mEnrich-seq: Methylation-guided enrichment sequencing of bacterial taxa of

interest from microbiome. *Nature Methods*. 2024;21:236–246 **https://doi.org/10.1038/s41592-023-02125-1**.

Casadesús and Low, 2006 Casadesús J, Low D. Epigenetic gene regulation in the bacterial world. *Microbiology and MolecularBiology Reviews*. 2006;70(3):830–856 **https://doi.org/10.1128/MMBR.00016-06**.

Charalampous et al., 2023 Charalampous T, Alcolea-Medina A, Snell LB, et al.... Routine metagenomics service for ICU patients with respiratory infection. *American Journal of Respiratory and Critical Care Medicine*. 2023;209(2):164–174 **https://doi.org/10.1164/rccm.202305-0901OC**.

Charalampous et al., 2019 Charalampous T, Kay GL, Richardson H, et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nature Biotechnology*. 2019;37:769–778 **https://doi.org/10.1038/s41587-019-0156-5**.

Cuscó et al., 2022 Cuscó A, Pérez D, Viñes J, Fàbregas N, Francino O. Novel canine high-quality metagenome-assembled genomes, prophages and host-associated plasmids provided by long-read metagenomics together with Hi-C proximity ligation. *Microbial Genomics*. 2022;8(3):000802 **https://doi.org/10.1099/mgen.0.000802**.

Dickson, 2021 Dickson I. *Sequencing the unculturable majority*. *Nature Portfolio* 2021; **https://www.nature.com/articles/d42859-020-00102-8** [Accessed 9 December 2024].

Durazzi et al., 2021 Durazzi F, Sala C, Castellani G, Manfreda G, Remondini D, De Cesare A. Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. *Scientific Reports*. 2021;11 **https://doi.org/10.1038/s41598-021-82726-y**.

Gowers et al., 2019 Gowers G-OF, Vince O, Charles J-H, Klarenberg I, Ellis T, Edwards A. Entirely off-grid and solar-powered DNA sequencing of microbial communities during an ice cap traverse expedition. *Genes*. 2019;10(11):902 **https://doi.org/10.3390/genes10110902**.

Haan and Drown, 2021 Haan TJ, Drown DM. Unearthing antibiotic resistance associated with disturbance-induced permafrost thaw in interior Alaska. *Microorganisms*. 2021;9(1):116 **https://doi.org/10.3390/microorganisms9010116**.

Hall et al., 2024 Hall MB, Wick RR, Judd LM, et al. Benchmarking reveals superiority of deep learning variant callers on bacterial nanopore sequence data. *eLife*. 2024;13:RP98300 **https://doi.org/10.7554/eLife.98300.3**.

Heidelbach et al., 2024 Heidelbach S, Dall SM, Bøjer JS, et al. Nanomotif: Identification and exploitation of DNA methylation motifs in metagenomes using Oxford Nanopore sequencing. *bioRXiv* 2024; **https://doi.org/10.1101/2024.04.29.591623**.

Kleikamp et al., 2023 Kleikamp HBC, Grouzdev D, Schaasberg P, et al. Metaproteomics, metagenomics and 16S rRNA sequencing provide different perspectives on the aerobic granular sludge microbiome. *Water Research*. 2023;246 **https://doi.org/10.1016/j.watres.2023.120700**.

Lerminiaux et al., 2024 Lerminiaux N, Fakharuddin K, Mulvey MR, Mataseje L. Do we still need Illumina sequencing data? Evaluating Oxford Nanopore Technologies R10.4.1 flow cells and the Rapid v14 library prep kit for Gram negative bacteria whole genome assemblies. *Canadian Journal of Microbiology*. 2024;70(5):178–189 **https://doi.org/10.1139/cjm-2023-0175**.

Liu et al., 2022 Liu L, Yang Y, Deng Y, Zhang T. Nanopore long-read-only metagenomics enables complete and high-quality genome

reconstruction from mock and complex metagenomes. *Microbiome*. 2022;10 **https://doi.org/10.1186/s40168-022-01415-8**.

Minor et al., 2023 Minor NR, Ramuta MD, Stauss MR, et al. Metagenomic sequencing detects human respiratory and enteric viruses in air samples collected from congregate settings. *Scientific Reports*. 2023;13:21398 **https://doi.org/10.1038/s41598-023-48352-6**.

Nicholls et al., 2019 Nicholls SM, Quick JC, Tang S, Loman NJ. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *GigaScience*. 2019;8(5):giz043 **https://doi.org/10.1093/gigascience/giz043**.

Oxford Nanopore Technologies, 2024a Oxford Nanopore Technologies. (2024a). Platform solution. **https://nanoporetech.com/platform** [Accessed 9 December 2024].

Oxford Nanopore Technologies, 2024b Oxford Nanopore Technologies. (2024b). Ultra-Long DNA Sequencing Kit. **https://store.nanoporetech.com/uk/ultra-long-dna-sequencing-kit-v14.html** [Accessed 9 December 2024].

Oxford Nanopore Technologies, 2024c Oxford Nanopore Technologies. (2024c). Research with impact at ASM Microbe. **https://nanoporetech.com/blog/research-with-impact-at-asm-microbe** [Accessed 9 December 2024].

Oxford Nanopore Technologies, 2024d Oxford Nanopore Technologies. (2024d). Sequence. **https://nanoporetech.com/products/sequence** [Accessed 10 December 2024].

Portik et al., 2022 Portik DM, Brown CT, Pierce-Ward NT. Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets. *BMC*

Bioinformatics. 2022;23 **https://doi.org/10.1186/s12859-022-05103-0**.

Pugh, 2023 Pugh J. The current state of nanopore sequencing. In: Arakawa K, ed. *Nanopore sequencing, methods and protocols*. New York: Humana; 2023;:3–14. **https://doi.org/10.1007/978-1-0716-2996-3_1**.

Reska et al., 2023 Reska T, Pozdniakova S, Borràs S, et al. Air monitoring by nanopore sequencing. *ISME Communications*. 2023;4(1):ycae099 **https://doi.org/10.1093/ismeco/ycae099**.

Sanderson et al., 2024 Sanderson ND, Hopkins KMV, Colpus M, et al. Evaluation of the accuracy of bacterial genome reconstruction with Oxford Nanopore R10.4.1 long-read-only sequencing. *Microbial Genomics*. 2024;10 **https://doi.org/10.1099/mgen.0.001246**.

Sauerborn et al., 2024 Sauerborn E, Corredor NC, Reska T, et al. Detection of hidden antibiotic resistance through real-time genomics. *Nature Communications*. 2024;15 **https://doi.org/10.1038/s41467-024-49851-4**.

Seitz et al., 2021 Seitz TJ, Schütte UME, Drown DM. Soil disturbance affects plant productivity via soil microbial community shifts. *Frontiers in Microbiology*. 2021;12 **https://doi.org/10.3389/fmicb.2021.619711**.

Sereika et al., 2022 Sereika M, Kirkegaard RH, Karst SM, et al. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nature Methods*. 2022;19:823–826 **https://doi.org/10.1038/s41592-022-01539-7**.

Shi et al., 2022 Shi Y, Wang G, Lau HC-H, Yu J. Metagenomic sequencing for microbial DNA in human samples: Emerging

technological advances. *International Journal of Molecular Sciences*. 2022;23(4):2181 **https://doi.org/10.3390/ijms23042181**.

Simon et al., 2023 Simon SA, Schmidt K, Griesdorn L, Soares AR, Bornemann TLV, Probst AJ. Dancing the Nanopore limbo—nanopore metagenomics from small DNA quantities for bacterial genome reconstruction. *BMC Genomics*. 2023;24 **https://doi.org/10.1186/s12864-023-09853-w**.

Urban et al., 2023 Urban L, Perlas A, Francino O, et al.... Real-time genomics for One Health. *Molecular Systems Biology*. 2023;19(8):e11686 **https://doi.org/10.15252/msb.202311686**.

Wilbanks et al., 2022 Wilbanks EG, Doré H, Ashby MH, Heiner C, Roberts RJ, Eisen JA. Metagenomic methylation patterns resolve bacterial genomes of unusual size and structural complexity. *The ISME Journal*. 2022;16(8):1921–1931 **https://doi.org/10.1038/s41396-022-01242-7**.

World Health Organization, 2024 World Health Organization. (2024). The top 10 causes of death. **https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death** [Accessed 10 December 2024].

Zhang et al., 2023 Zhang T, Li H, Ma S, et al. The newest Oxford Nanopore R10.4.1 full-length 16S rRNA sequencing enables the accurate resolution of species-level microbial community profiling. *Applied and Environmental Microbiology*. 2023;89:e00605–23 **https://doi.org/10.1128/aem.00605-23**.

Zhao et al., 2023 Zhao W, Zeng W, Pang B, et al. Oxford nanopore long-read sequencing enables the generation of complete bacterial and plasmid genomes without short-read sequencing.

Frontiers in Microbiology. 2023;14
https://doi.org/10.3389/fmicb.2023.1179966.

CHAPTER 8

Highly accurate long-read PacBio HiFi sequencing for microbiome and metagenomic applications

Jeremy E. Wilkinson and Daniel M. Portik, Pacific Biosciences of California, Inc., Menlo Park, CA, United States

Abstract

In recent years long-read sequencing has transformed the way we approah microbiome sequencing applications. In this chapter we focus on the impact of PacBio HiFi sequencing on targeted amplicon sequencing (16 S), taxonomic profiling of shotgun metagenomic reads, and metagenome assembly. For targeted amplicon sequencing full-length 16 S or internal transcribed spacer (ITS) sequencing vastly outperforms partial 16 S/ITS sequencing, and the accuracy of HiFi reads leads to better taxonomic assignments. In a metagenomic context HiFi reads improve taxonomic classification and profiling, and newly developed long-read methods minimize false positives while detecting nearly all species present. Finally HiFi sequencing has revolutionized metagenome assembly, and large numbers of high-quality metagenome-assembled genomes (MAGs) can now be recovered across a variety of host-associated and environmental microbiomes. As innovations in HiFi sequencing continue to improve accessibility, HiFi reads are expected to play a key role in advancing microbiome and metagenomic studies.

Keywords

Metagenome-assembled genomes; long-read sequencing; taxonomic precision; HiFi read; 16S rRNA; metagenome profiling

Introduction

Microorganisms are crucial to the ecosystems that support human, plant, and animal life, and their environments. They engage in various metabolic activities, produce and break down compounds, and significantly influence health, fitness, phenotypes, and ecological dynamics. The human microbiome, in particular, has been shown to be vital for health or disease, with imbalances in the gut microbiome linked to conditions such as inflammatory bowel disease, diabetes, cardiovascular disease, colon cancer, and neurological disorders (Lynch & Pedersen, 2016; Wang & Jia, 2016; Duvallet et al., 2017; Gilbert et al., 2018; Lloyd-Price et al., 2019). Likewise soil and water microbes actively contribute to the health of organisms, agriculture, and ecosystems (Banerjee & van der Heijden, 2023). These connections highlight the growing importance of thoroughly characterizing microbiome species and linking them to biological outcomes. Metagenomics and microbiome research are playing an increasingly important role in fields like public health, epidemiology, and therapeutics. However there is still a great deal of work to be performed and many discoveries yet to be made (Wilkinson et al., 2021).

Despite considerable efforts to characterize the microbial diversity present in the human gut substantial numbers of microorganisms still remain uncultured and unknown (Lagier et al., 2018; Almeida et al., 2019; Forster et al., 2019; Nayfach et al., 2019; Pasolli et al., 2019; Zou et al., 2019; Almeida et al., 2021), and this shortcoming is even more apparent in complex environmental samples such as soil, water, and sediment (Vitorino & Bessa, 2018; Baldrian, 2019; Anthony et al., 2024). The characterization of microbial diversity has historically Page 164 occurred through isolation and culturing methods, but recently the pace of discovery has accelerated using metagenomic sequencing. Recovering high-quality genomes through this

approach has been historically challenging using short-read sequencing. For the task of detecting microbes both targeted (e.g., 16 S rRNA sequencing) and nontargeted (e.g., metagenomic read classification) approaches are commonly used. Targeted amplicon sequencing (e.g., 16 S) is an economical method for evaluating the composition of microbial communities, particularly in low-biomass samples. However the success of both targeted and nontargeted detection approaches relies on databases built from reliable microbial genomes, and capturing new diversity is therefore an essential task.

In the last few years long-read sequencing has been transformative for targeted amplicon and metagenomic sequencing experiments (Agustinho et al., 2024). The two prevalent long-read sequencing technologies are Pacific Biosciences® (PacBio®) and Oxford Nanopore Technologies (ONT). PacBio employs single-molecule, real-time (SMRT®) sequencing, which produces long and highly accurate consensus reads that are referred to as HiFi reads. The HiFi reads are up to ~25 kb in length and minimally Q20 in quality (representing 99% accuracy) but often display a median of Q30 or 99.9% accuracy (Wenger et al., 2019). As of December 2024 there are currently three PacBio HiFi long-read sequencers available: the Sequel® II, Revio®, and Vega™ systems. The Sequel II system was the first to employ HiFi sequencing, and it uses the 8 M SMRT® Cell, can run for 10–30 hours, and produces 2–4 million HiFi reads with up to 40 Gbp of HiFi data. The Revio system was released as a higher-throughput instrument, and it can run four SMRT cells in parallel using 12-, 24-, or 30-hour run times. Each Revio SMRT cell with SPRQ™ chemistry can generate 7–14 million HiFi reads and up to 120 Gbp of HiFi data, for a combined total of up to 480 Gbp in 24 hours. The Vega system is a new benchtop sequencer that can run one SMRT cell at a time for 12- or 24-hour run times and at the time of release is capable of producing 4–8 million HiFi reads and 60 Gbp of HiFi data.

Here we describe the impact of PacBio HiFi sequencing on three applications related to microbiome and metagenomic studies. These include full-length gene amplicon sequencing, taxonomic and functional profiling of

metagenomic reads, and metagenome assembly. We describe how long-read sequencing has impacted each of these applications, highlighting major advances in approaches and methods. Page 165

Full-length gene amplicon microbiome sequencing

Targeted amplicon sequencing of a single gene (e.g., 16 S rRNA) is a cost-effective way of assessing the composition of metagenomic samples. The 16 S rRNA gene is a highly conserved gene found among all bacteria and contains nine hypervariable regions which can be used for clade identification.

However, due to the high similarity in 16 S rRNA genes in closely related bacteria, it is crucial to sequence the entire 16 S gene (V1–V9, ~1.5 kb) and with high accuracy to enable precise species- or strain-level identification. Several comparative studies have demonstrated that PacBio HiFi full-length (FL) 16 S sequencing (covering V1–V9 regions) offers greater taxonomic resolution and a higher proportion of classified reads as compared to next-generation sequencing (NGS) short-read 16 S sequencing, which covers only 1–3 variable regions of the 16 S gene (Johnson et al., 2019; Lefoulon et al., 2021; Notario et al., 2023; Buetas et al., 2024; Chuang et al., 2024; Gao et al., 2024). Among these comparative studies Notario et al. (2023) compared three different methods to determine which method achieves the best taxonomic resolution using 16 S rRNA sequencing: (1) a single- or two-variable region approach, (2) a multiplex approach, covering several regions of the 16 S rRNA gene with both of the first two approaches using NGS short-read sequencing, and (3) a full-length approach, which analyzes the entire length of the 16 S rRNA gene with PacBio HiFi long-read sequencing. They found the full-length 16 S rRNA approach yielded the greatest discriminating power up to the species level. It also benefited from less labor, less time, and lower contamination risk, at a similar cost to the standard single- and two-variable region approach (**Fig. 8.1**). Therefore full-length 16 S rRNA sequencing with PacBio HiFi long reads demonstrated an improvement in experimental

workflow, data accuracy, and taxonomic resolution compared with other standard methods.

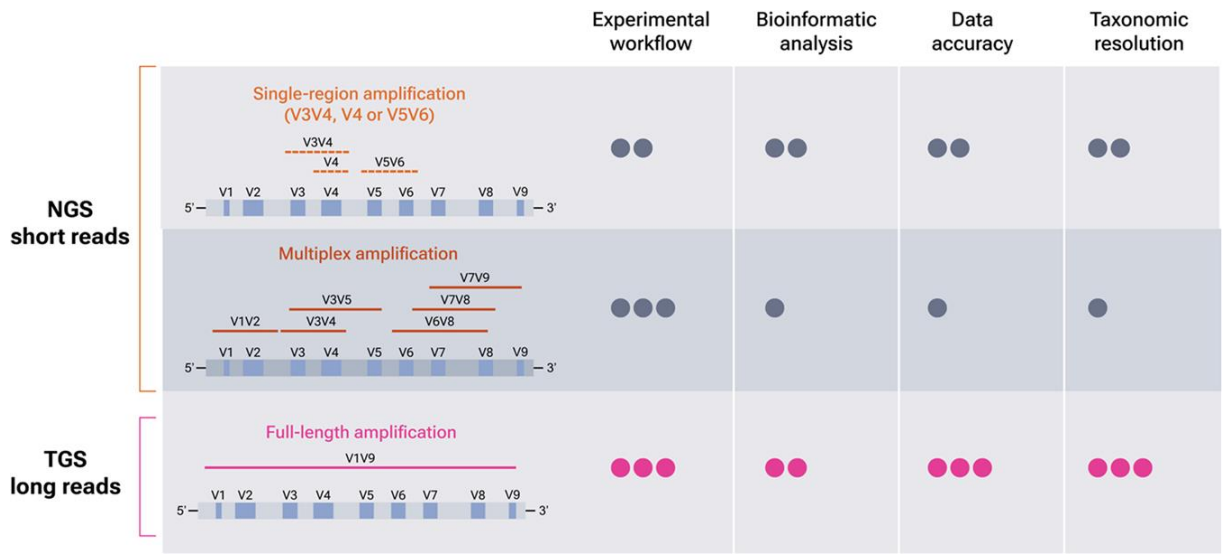


FIGURE 8.1 Comparison results of three different methods for 16 S rRNA amplicon sequencing. Full-length 16 S rRNA sequencing with PacBio technology covers the entire 1.5 kb 16 S rRNA gene with high accuracy, allowing for species- and strain-level identification in microbiome studies and outcompetes the two short-read methods: single-region or two-region partial 16 S, and multiplex amplification. Adapted from Notario et al. (2023).

Follow for extended description

Taxonomic precision is also essential for community ecology studies involving microbial eukaryotes. For organisms like fungi, which are challenging to culture, the taxonomic precision offered by culture-free methods enhances our understanding of their roles within the mycobiome. Molecular classification approaches rely on sequencing specific target regions selected for their evolutionary rates and ease of amplification and detection. The most used barcode sequences for eukaryotes include the internal transcribed spacer regions (ITS1 and ITS2) and the 18 S and 28 S rRNA genes. PacBio HiFi long-read amplicon sequencing Page 166 also has significant

advantages for high-throughput, high-resolution eukaryotic barcode sequencing. Unlike short-read methods, which only cover the ITS1 or ITS2 regions, HiFi reads span the entire length of fungal ITS regions, typically around 800 bp, with the potential for amplicons up to ~5 kb if the 18 S and/or 28 S genes are included to cover the entire operon. This longer amplicon approach, which includes more conserved rRNA genes in addition to the ITS regions, reduces sequence similarity and provides higher-resolution taxonomic information at the species level (Tedersoo et al., 2018; Tedersoo et al., 2021; Latz et al., 2022; Runnel et al., 2022). Although the length of the reads is important, accuracy is still critical and HiFi sequencing displays higher precision than other long-read approaches (Overgaard et al., 2024).

In addition to delivering accurate taxonomic assignments HiFi full-length amplicon sequencing has become cost-comparable to short-read, partial amplicon sequencing. The PacBio Kinnex™ 16 S rRNA kit uses amplified 16 S amplicons from DNA as input and produces a sequencing-ready library that enhances throughput by up to 12-fold compared with standard FL 16 S libraries. This kit is based on the multiplexed array sequencing (MAS-Seq) method (Al'Khafaji et al., 2024) which is applied to FL 16 S amplicons. The MAS-Seq method is a versatile throughput increase method that takes advantage of the longer HiFi read lengths to concatenate smaller amplicons into ordered arrays with programmable array sizes via parallel PCRs, unique barcodes, and Page 167 specific adapters for digestion and ligation. This concatenation approach results in significantly higher-throughput FL 16 S sequencing, with the capability to multiplex up to 1536 amplicon samples per SMRT cell (**Fig. 8.2**). The Kinnex 16 S rRNA kit was tested on a diverse range of sample types including mock community standards, feces, saliva, plant, soil, wastewater, and various swabs (skin, oral, vaginal, and veterinary wound). Sequencing data were analyzed using the HiFi-16S-workflow (available on GitHub), which provides a FASTQ-to-report analysis solution for FL 16 S HiFi reads.

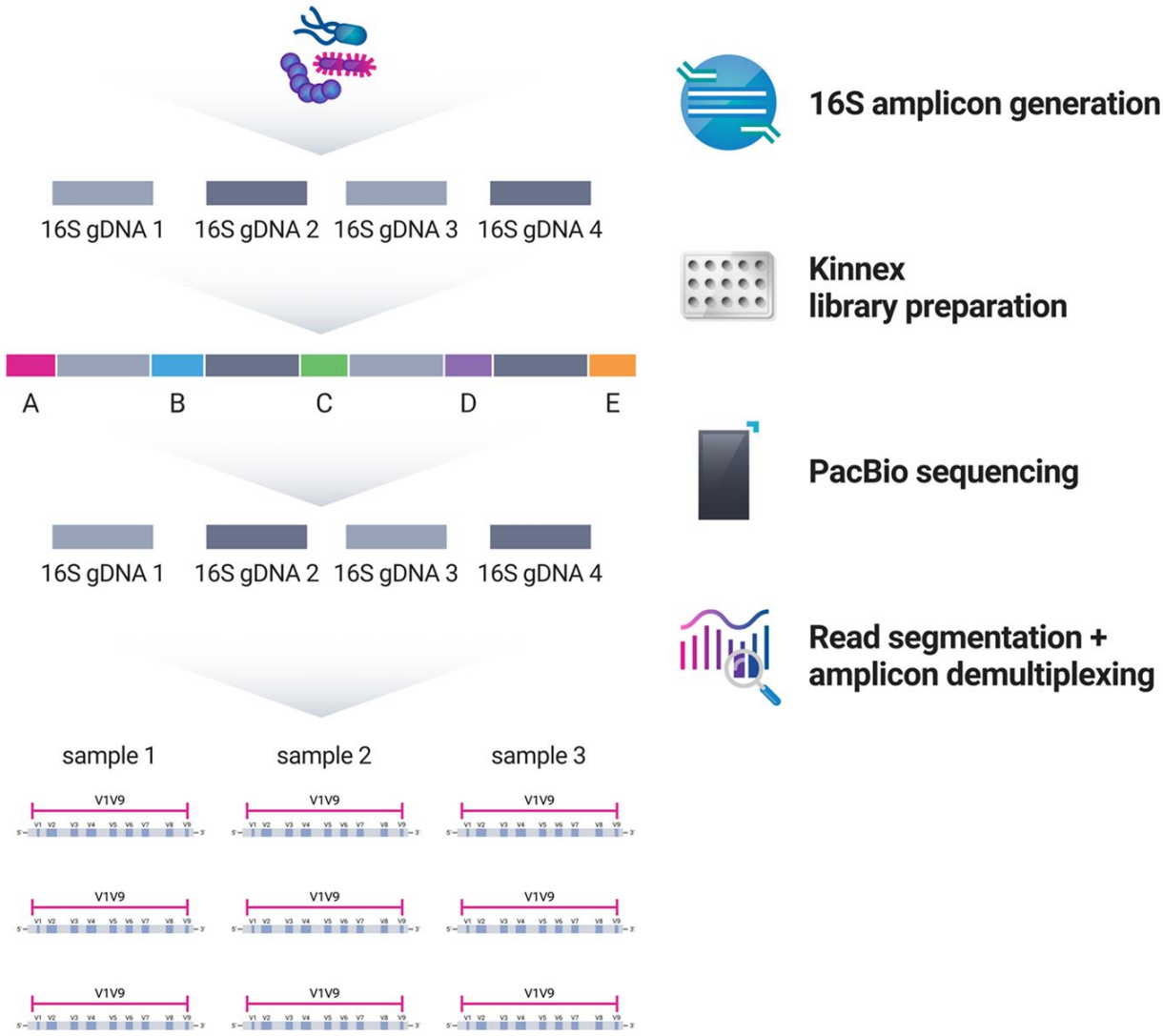


FIGURE 8.2 Kinnex 16 S rRNA sequencing workflow. A pooled library of barcoded full-length 16 S amplicons from samples is concatenated into a large-insert library with the Kinnex 16 S rRNA kit that uses 12 concatemers, sequenced on a PacBio HiFi long-read system, and then processed using PacBio SMRT Link software.

Follow for extended description

The results show that Kinnex 16 S sequencing can yield an average of more than 30,000 reads per sample, based on a 1536-plex on a single Revio SMRT cell or at a 768-plex on a Sequel IIe SMRT cell 8 M (**Fig. 8.3**). A standard FL 16 S library on the Sequel II systems achieves Page 168 2.1–3.3 million (M) reads, but with Kinnex there are a total of 19.6–28.4 M FL 16 S reads (**Fig.**

8.3A). The difference is also apparent on the Revio system, with 8.5–10.1 M reads for standard FL 16 S sequencing and 62.5–72.2 M FL 16 S reads with Kinnex. We compared the taxonomic classifications of Kinnex 16 S to those from standard FL 16 S and found a significant correlation in species composition (**Fig. 8.3B**). For a mock community composed of 20 species (ATCC MSA-1003) the Kinnex FL 16 S results provided a more accurate representation of the expected species abundances due to higher read depth (**Fig. 8.3C**).

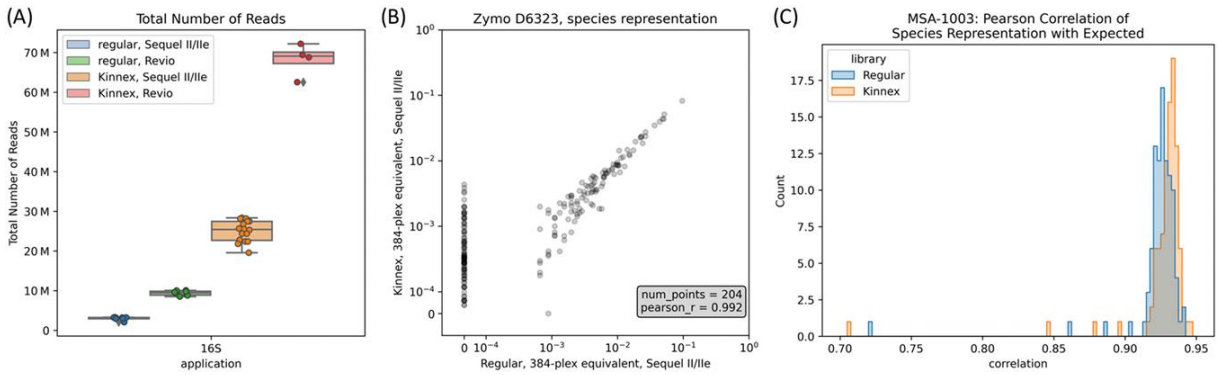


FIGURE 8.3 Kinnex 16 S rRNA kit results compared with standard FL 16 S. (A) Kinnex concatenation results in an up to 12-fold throughput increase for full-length 16 S sequencing on Sequel II/IIe and Revio systems. Each datapoint in the box plot represents a regular or Kinnex 16 S library run on one SMRT cell. The number of reads is the de-concatenated S-reads representing the original full-length 16 S amplicon. Outliers are indicated by diamonds. (B) Kinnex 16 S is highly correlated to standard FL 16 S for species composition. 384-plex of the ZymoBIOMICS fecal reference sample (D6323) on both Kinnex 16 S and regular FL 16 S on the Sequel IIe system. Pearson correlation is calculated on the species relative abundances between the two datasets. (C) Kinnex allows higher multiplexing and captures better species representation in microbiome standard. 96-plex of the same ATCC MSA-1003 sample for both Kinnex 16 S and regular FL 16 S. Pearson correlation is calculated based on the fraction of reads assigned to species against the expected fraction.

Follow for extended description

Taxonomic and functional profiling of metagenomic reads

Shotgun metagenomics aims to sequence the genomic DNA of all organisms in a microbiome, and the resulting sequencing reads can be used to predict the identities and abundances of microbial species. There has been extensive development of taxonomic profiling and classification methods for short-read technologies [reviewed by Breitwieser et al. (2018)]; however this approach is fundamentally limited by the small volume of data contained per short read (e.g., 50–250 bp). The size distribution of long-read datasets can vary across metagenomic samples, but the mean read length of published HiFi metagenomic datasets ranges from 7 to 19 kb (see **Table 8.1**). This means that individual HiFi reads Page 169 Page 170 Page 171 typically span multiple genes and intergenic regions, and this larger volume of information could be used to improve the reliability of both taxonomic and functional annotations.

Table 8.1

Type	Sample	Instrument	SMRT cells	HiFi reads (million)	Mean read length (kb)	Total data (Gb)	Mean QV	Total MAGs	HQ-MAGs	Source	Accession
Artificial	Digester/Bioreactor	Sequel II	1 (8 M)	0.99	15.4	15.3	Q31	168	83	**Sereika et al. (2022)**	ERR7015089
	Digester/Bioreactor	Sequel II	1 (8 M)	3.15	8.5	26.9	Q29	123	72	**Wang et al. (2024)**	SRR24881069
Environmental	Lichen thallus	Sequel II	1 (8 M)	3.58	6.8	24.7	Q42	46	25	(**Ahmad et al., 2023**)	SRR24475747
	Seawater	Sequel II	1 (8 M)	2.92	7.7	22.4	Q39	121	53	**Orellana et al. (2023)**	ERR9769281
	Seawater	Sequel II	1 (8 M)	2.55	8.1	20.6	Q37	123	41	**Orellana et al. (2023)**	ERR9769303
	Hot spring sediment	Sequel II	1 (8 M)	2.69	10.3	27.9	Q31	121	86	**Kato et al. (2022)**	DRR290133
	Wetland soil	Revio	3 (25 M)	20.64	8.4	195.8	Q41	1232	563	Currently unpublished	–
	Grassland soil 1	Revio	4 (25 M)	30.56	8.7	283.2	Q40	397	133	Currently unpublished	–
	Grassland soil 2	Revio	4 (25 M)	17.38	8.2	151.1	Q42	412	206	Currently unpublished	–
Gut microbiome	Sheep gut	Sequel II	8 (8 M)	18.46	11.2	206.5	Q35	1036	579	**Bickhart et al. (2022)**	SRR14289618
	Chicken gut	Sequel II	2 (8 M)	5.87	18.9	111.9	Q31	552	269	**Zhang et al. (2023)**	SRR19732729
	Human gut	Sequel II	1 (8 M)	1.91	9.8	18.8	Q39	211	90	**Kim et al. (2022)**	SRR15275211
	Human gut	Sequel II	1 (8 M)	1.79	10.3	18.5	Q40	179	85	**Kim et al. (2022)**	SRR15275213
	Human gut (Zymo D6323)	Sequel II	8 (8 M)	17.15	7.1	120.8	Q40	362	174	**Portik et al. (2024)**	SRR28996636
	Human gut (Zymo D6323)	Revio	2 (25 M)	17.56	7.7	135.1	Q45	378	177	**Portik et al. (2024)**	SRR28996637
	Human gut (Zymo D6323)	Combined	(see above)	34.71	7.5	255.9	Q42	480	239	**Portik et al. (2024)**	(see above)

Despite the potential of long reads to produce more accurate taxonomic classification and profiling results, until recently there were few taxonomic methods available for long reads. Among the first to appear were MEGAN-LR (Huson et al., 2018), MetaMaps (Dilthey et al., 2019), CDKAM (Bui & Wei, 2020), BugSeq (Fan et al., 2021), and MMseqs2 taxonomy (Mirdita et al., 2021). Additionally sourmash (Brown & Irber, 2016; Pierce et al., 2019; Irber et al., 2022) was extended to perform taxonomic profiling using long- or short-read metagenomic data. Portik et al. (2022) conducted the first long-read benchmarking study to examine the performance of these methods, along with several popular short-read taxonomic methods. Their study included PacBio, ONT, and Illumina sequencing datasets for mock

communities and compared precision and recall among other metrics. The study revealed several key findings, including (1) long-read datasets achieved substantially higher precision relative to the short-read datasets, (2) methods designed for short reads resulted in high false positives and consequently low precision when used with long reads, (3) threshold filtering can increase precision in underperforming methods but it lowers recall, and (4) sourmash, MEGAN-LR, and BugSeq were the top-performing methods for HiFi data (achieving high or perfect precision with high recall). Importantly, the pattern of high-precision and high recall was not achievable using short reads, despite applying various levels of threshold filtering. Finally the high accuracy of HiFi reads worked better with methods utilizing kmers or protein prediction (sourmash, DIAMOND & MEGAN-LR, MMseqs2), resulting in a higher proportion of classified reads than the ONT datasets. A more recent benchmarking study tested some of the same methods and found overall similar patterns (Mari et al., 2024). Based on their work, Portik et al. (2022) proposed sourmash is the best overall profiling tool for HiFi data, followed by DIAMOND & MEGAN-LR and BugSeq as the best classifiers (using protein and nucleotide alignments, respectively).

Since the evaluation performed by Portik et al. (2022) there has been a rapid increase in the development of methods that are compatible with long reads. Over two years (2023–2024), at least nine new classifiers or profilers have appeared (**Fig. 8.4**). In chronological order these include Metabuli (Kim & Steinegger, 2024), Taxor (Ulrich & Renard, 2024), Page 172 sylph (Shaw & Yu, 2024), MetageNN (da Silva et al., 2024), Melon (Chen et al., 2024), CAIM (Acheampong et al., 2024), Mibianto (Hirsch et al., 2024), Lemur/Magnet (Sapoval et al., 2024), and PanTax (Zhang et al., 2024). Each of these methods is likely to have unique strengths and weaknesses and potentially different use cases. However a comprehensive evaluation of these methods is currently lacking. It is not clear how they perform relative to one another and to the methods identified in previous benchmarks. A new comparative benchmarking study would be highly beneficial, particularly as

the number of available tools continues to grow. Overall, the rapid development of new taxonomic profiling and classification methods indicates that the research community is increasingly recognizing the benefits of long reads.

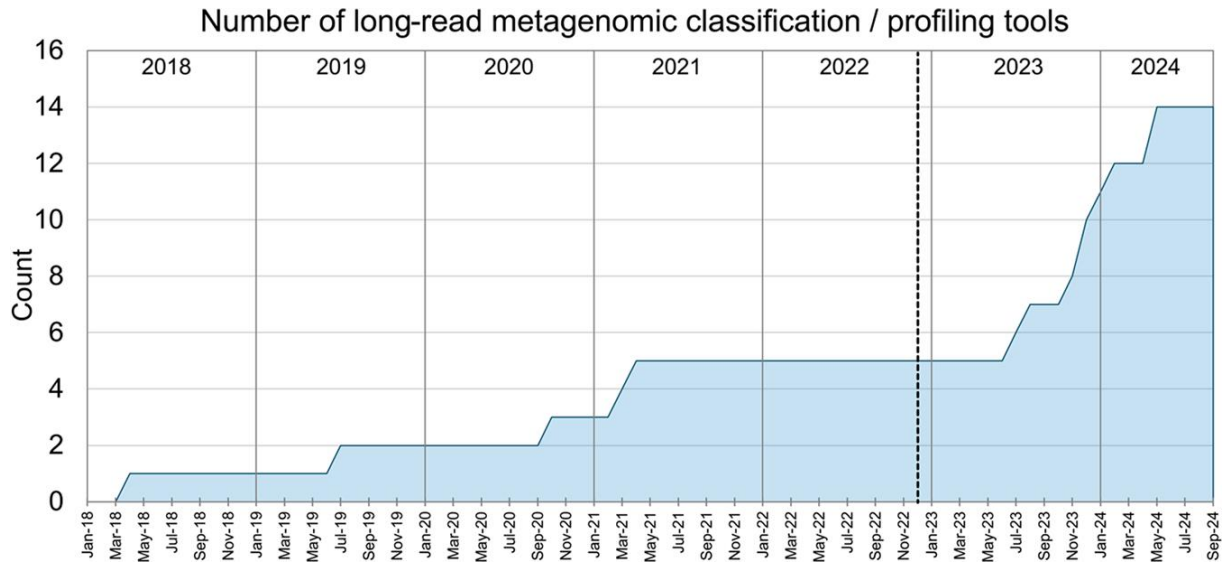


FIGURE 8.4 The accumulation of taxonomic profiling and classification methods designed for long reads, over time. The dotted line represents the publication of the first long-read benchmarking study, conducted by Portik et al. (2022).

Follow for extended description

The presence of multiple, complete genes on each HiFi read offers the opportunity to identify and quantify genes based on functions, pathways, biosynthetic gene clusters (BGCs), antimicrobial resistance (AMR), and other classification categories. Several taxonomic classification methods rely on the identification of ORFs and proteins in long reads, including MEGAN-LR, MMseqs2, and Metabuli. However only MEGAN-LR currently cross-references protein alignments with various databases such as KEGG, InterPro2Go, SEED, EC, and eggNOG to provide annotations per read (Huson et al., 2018). Taken together the annotations aggregated across reads provide

functional profiles for samples, in addition to taxonomic profiles. Using this methodology for a gut microbiome study Gehrig et al. (2022) found Page 173 that an average of 86% of HiFi reads received at least one functional annotation, and that the average number of annotations per read was four (for average read lengths of 5–7 kb). Furthermore, many of these annotations represented complete gene regions. In a different study, Orellana et al. (2023) performed gene prediction in metagenomic HiFi reads from seawater, comparing the strict translation method of Prodigal (Hyatt et al., 2010) to the indel-aware approach implemented in FragGeneScan (Rho et al., 2010). They determined that FragGeneScan produced more accurate protein sequences from the HiFi reads. Although they gathered a large set of predicted proteins, they did not attempt to classify them. Most recently Yadav & Subramanian (2024) developed HiFiBGC, a workflow for BGC detection in HiFi metagenomes. HiFiBGC assembles contigs using multiple methods, runs antiSMASH (Blin et al., 2023) on the contigs and unassembled reads, and de-replicates resulting BGCs through clustering approaches. Up to 78% more BGCs can be recovered through this method versus other approaches, which is attributed to better quality assemblies and the ability to identify BGCs directly in the unassembled reads (Yadav & Subramanian, 2024). Together these studies demonstrate that identifying genes and proteins in HiFi reads is straightforward using DIAMOND, FragGeneScan, antiSMASH, or other methods. However, the read-based analysis of proteins remains an under-utilized aspect of HiFi sequencing. Creating new tools or workflows to streamline this process, including protein alignment or prediction, identification and classification, and aggregation would be a particularly useful direction for future work.

Metagenome assembly

The goal of metagenome assembly is to reconstruct complete genomes for all microorganisms present in a microbiome sample. Genomes obtained through metagenome assembly and binning are commonly referred to as

metagenome-assembled genomes, or MAGs. Metagenome assembly has been historically challenging using short-read technologies, because repetitive genomic regions and interspecies genomic overlaps typically result in shorter assembled contigs (Chen et al., 2020). Downstream binning steps, which attempt to group contigs belonging to the same putative genomes, can also introduce major errors. Therefore a large proportion of the resulting short-read MAGs consist of fragmented, error-prone genomes, and high-quality Page 174 single-contig MAGs are generally only possible through substantial manual curation efforts.

Long-read sequencing can vastly improve metagenome assembly because the longer read lengths typically span problematic difficult-to-resolve regions such as repeats or interspecies conserved regions (Kolmogorov et al., 2020; Feng et al., 2022; Agustinho et al., 2024; Benoit et al. 2024). Due to the complexity of metagenomic data (e.g., dozens to thousands of species and strains with variable abundances), an equally important consideration is base-pair level accuracy of the reads. Multiple metagenome assembly methods have been developed to leverage the higher accuracy of HiFi reads, including hifiasm-meta (Feng et al., 2022) and metaMDBG (Benoit et al., 2024). Performing metagenome assembly with HiFi reads routinely produces complete, single-contig MAGs (Bickhart et al., 2022; Feng et al., 2022; Kato et al., 2022; Kim et al., 2022; Jiang et al., 2023; Saak et al., 2023; Schaerer et al., 2023; Zhang et al., 2023; Benoit et al., 2024; Masuda et al., 2024; Minich et al., 2024; Portik et al., 2022), which are visible as circular contigs in assembly graphs (**Fig. 8.5**). Assemblies with HiFi reads generally produce more total MAGs and higher quality MAGs than short-read sequencing (Priest et al., 2021; Gehrig et al., 2022; Meslier et al., 2022; Orellana et al., 2023; Tao et al., 2023; Zhang et al., 2023; Eisenhofer et al., 2024; Minich et al., 2024). Similarly, because HiFi reads can be used with assembly algorithms designed for higher Page 175 accuracy, HiFi sequencing also currently outperforms ONT for metagenome assembly (Meslier et al., 2022; Sereika et al., 2022; Minich et al., 2024).

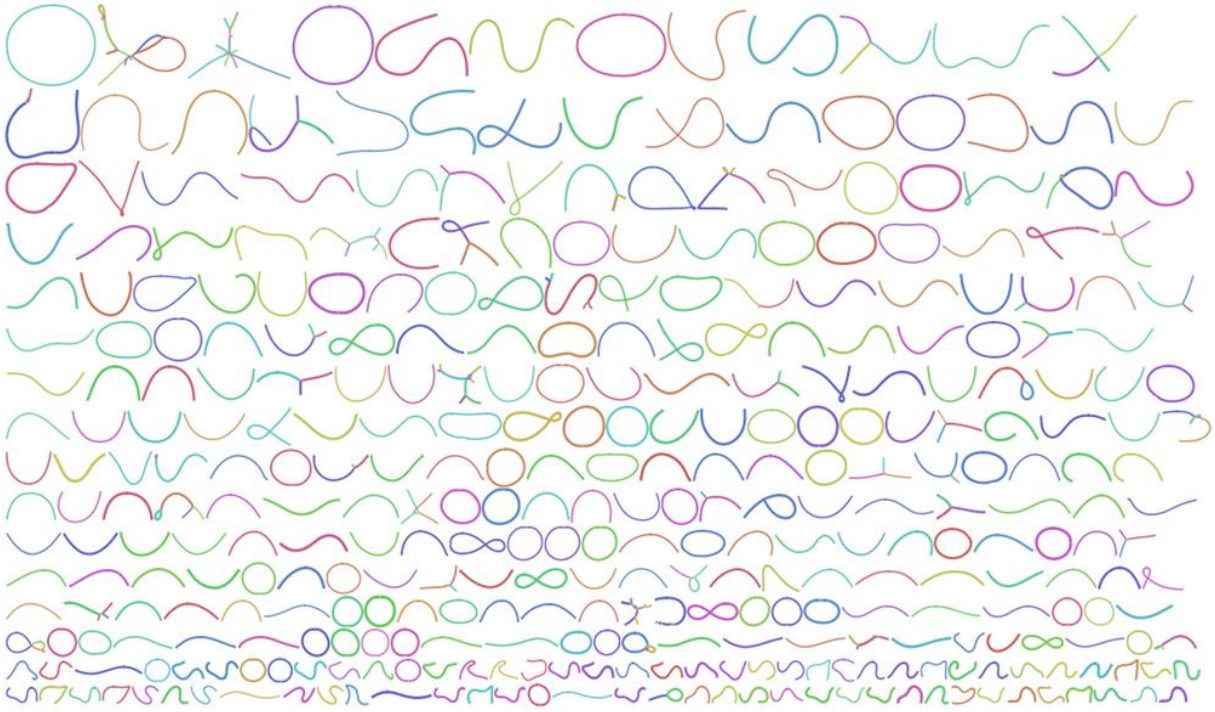


FIGURE 8.5 An example of a metagenome assembly graph. A partial assembly graph is shown for a commercially available pooled human fecal reference (ZymoBIOMICS D6323). The graph was created with hifiasm-meta. Lines with distinct colors represent individual contigs, and they are arranged by decreasing length. This dataset consisted of 34 million HiFi reads and 256 Gbp of data generated by a combination of runs on the Sequel IIe and Revio systems (**Table 8.1**). There were 98 complete, circular MAGs recovered in this particular assembly, ranging from 1.5 to 6 Mb in size.

Metagenome assembly with HiFi reads can produce complete MAGs, but many additional genomes will consist of two or more contigs. Capturing these MAGs consisting of multiple contigs requires a binning approach. Most binning methods were designed for short-read assemblies, and they perform clustering using tetranucleotide frequencies, depth of coverage, or deep learning methods (Sczyrba et al., 2017; Meyer et al., 2022). Using these methods to bin long-read assemblies can produce unexpected results, such as

the removal of complete, single-contig MAGs (Feng et al., 2022; Benoit et al., 2024). This can occur when a long and highly complete contig is mis-binned with additional contigs (representing strain variation or other sequence types). Generally this results in inflated contamination scores due to the presence of multiple copies of single-copy genes (SCGs) in the contig set, and the bin is removed because it fails to pass QC scores. Several new long-read binning algorithms have been developed, including MetaBCC-LR (Wickramarachchi et al., 2020), LRBinner (Wickramarachchi & Lin 2022), GraphMB (Lamurias et al., 2022), SemiBin2 (Pan et al., 2023), and TaxVAMB (Kutuzova et al., 2024). In particular SemiBin2 shows promising results for long-read metagenomic assemblies, producing up to 26% more bins than the next best method (Pan et al., 2023). In addition to these bioinformatic-based approaches proximity ligation information (e.g., Hi-C) can be used to guide binning (Burton et al., 2014; Press et al., 2017; Demaere & Darling, 2019). Proximity ligation crosslinks DNA strands that are physically close together and, in prokaryotes, allows downstream identification of regions of DNA that are close to one another within an individual cell (including mobile genetic elements). This spatial information can be used to guide or refine contig binning and link mobile genetic elements to their host chromosomes. The proximity ligation technology has been successfully paired with HiFi data in several studies (Bickhart et al., 2022; Saak et al., 2023; Portik et al., 2022). Many assembly workflows include multiple binning methods, and the results are de-replicated and merged to improve MAG quality and yield (Olm et al., 2017; Portik et al., 2024). After these steps MAGs often undergo filtration based on quality scores, followed by taxonomic annotation. The HiFi-MAG-Pipeline is an example of a workflow designed for long-read assemblies that includes binning, QC, and annotation steps (Portik et al., 2024).

 This workflow is part of the PacBio pb-metagenomics-tools package, available on GitHub. The workflow uses a completeness-aware binning strategy, which separates complete single-contig MAGs from the remaining incomplete contigs during the main binning steps. Binning of the

incomplete contigs is performed with SemiBin2 (Pan et al., 2023) and MetaBAT2 (Kang et al., 2019), and DAS_Tool (Sieber et al., 2018) is used for de-replication and merging. Bins are screened for quality based on user-defined thresholds for SCG completeness and contamination scores from CheckM2 (Chklovski et al., 2023). All filtered MAGs are then assigned to taxonomy using the Genome Taxonomy Database Toolkit (GTDB-Tk) v2.1.1 (Chaumeil et al., 2020; Chaumeil et al., 2022). Beyond the standard HiFi-MAG-Pipeline, a new HiFi metagenome assembly workflow called metaWGS was developed by Mainguy et al. (2024). The metaWGS workflow offers an alternative binning strategy (Binette) which produces similar or better results than the HiFi-MAG-Pipeline.

The combination of HiFi-specific assembly methods (hifiasm-meta, metaMDBG) and workflows designed to process long-read assemblies (HiFi-MAG-Pipeline) can be used to recover high-quality MAGs across a variety of metagenomic samples (**Fig. 8.6**). Previous studies focusing on HiFi metagenomics have used different combinations of assembly and binning methods to generate MAGs, making direct comparisons of MAG yields difficult. To allow direct comparisons across different sample types we performed an analysis of 12 publicly available HiFi datasets and three currently unpublished datasets (**Table 8.1**). The datasets represent a variety of microbiomes, including bioreactors and digesters, seawater, sediment, soils, sheep gut, chicken gut, and human gut. For all samples we used hifiasm-meta and metaMDBG to generate contigs, which were then analyzed using the HiFi-MAG-Pipeline separately (**Fig. 8.6)**. For total counts we included all MAGs displaying ≥50% SCG completeness and ≤10% contamination and labeled MAGs as high-quality (HQ-MAGs) if they displayed ≥90% SCG completeness and ≤5% contamination (Kyrpides et al., 2017). For each sample we report the MAG counts from the analysis with the highest number (e.g., either hifiasm-meta or metaMDBG). Finally there are at least four samples which utilize the Revio system for deep-sequencing (**Fig. 8.7**; 150–280 Gbp data): three soil samples and the combined dataset

for the pooled human gut (Sequel II & Revio systems; ZymoBIOMICS Fecal Reference D6323). The considerable amounts of total data for these samples allowed for an exploration of the relationship between total data and Page 177 MAG yields. We performed downsampling on these deeply sequenced samples to generate smaller datasets. We analyzed all downsampled datasets using the same methods described above. We then assessed if the number of MAGs could be predicted by the total data (Gb) per sample, either by a linear relationship (by performing linear regression) or a nonlinear relationship such as a saturation curve (by log-transforming the total data values before linear regression). A linear relationship suggests that additional sequencing would facilitate recovery of additional genomes, whereas a saturation curve suggests that most genomes are recovered in the sample.

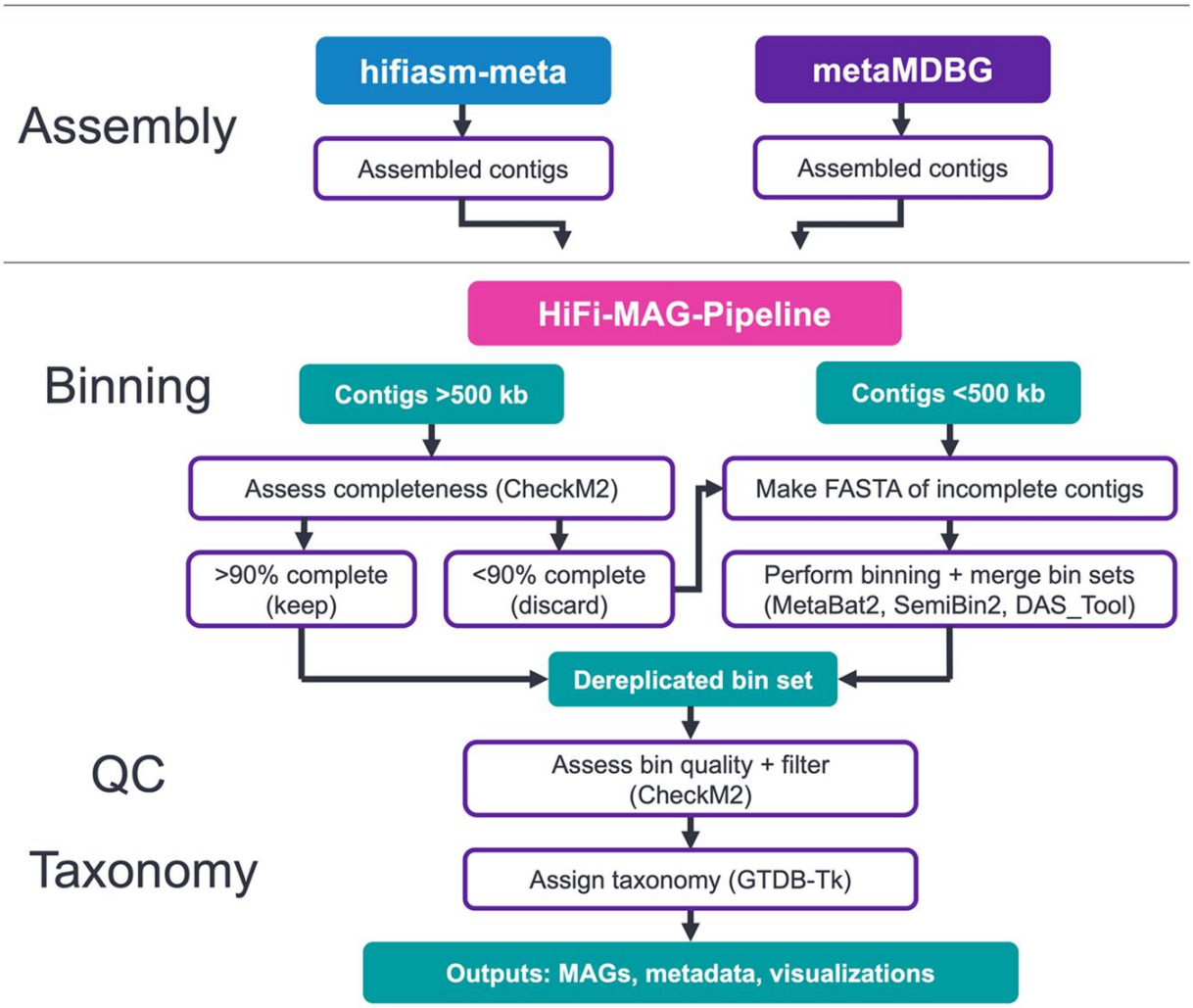


FIGURE 8.6 An overview of the current best practices for metagenome assembly and binning of HiFi metagenomic datasets. The assembly step consists of using hifiasm-meta or metaMDBG (or both) to generate contig sets. A given contig set is used as the input for the HiFi-MAG-Pipeline, which performs binning, QC, and taxonomic annotation steps. The workflow outputs MAG sequences, MAG metadata, and visualizations related to MAG quality.

Follow for extended description

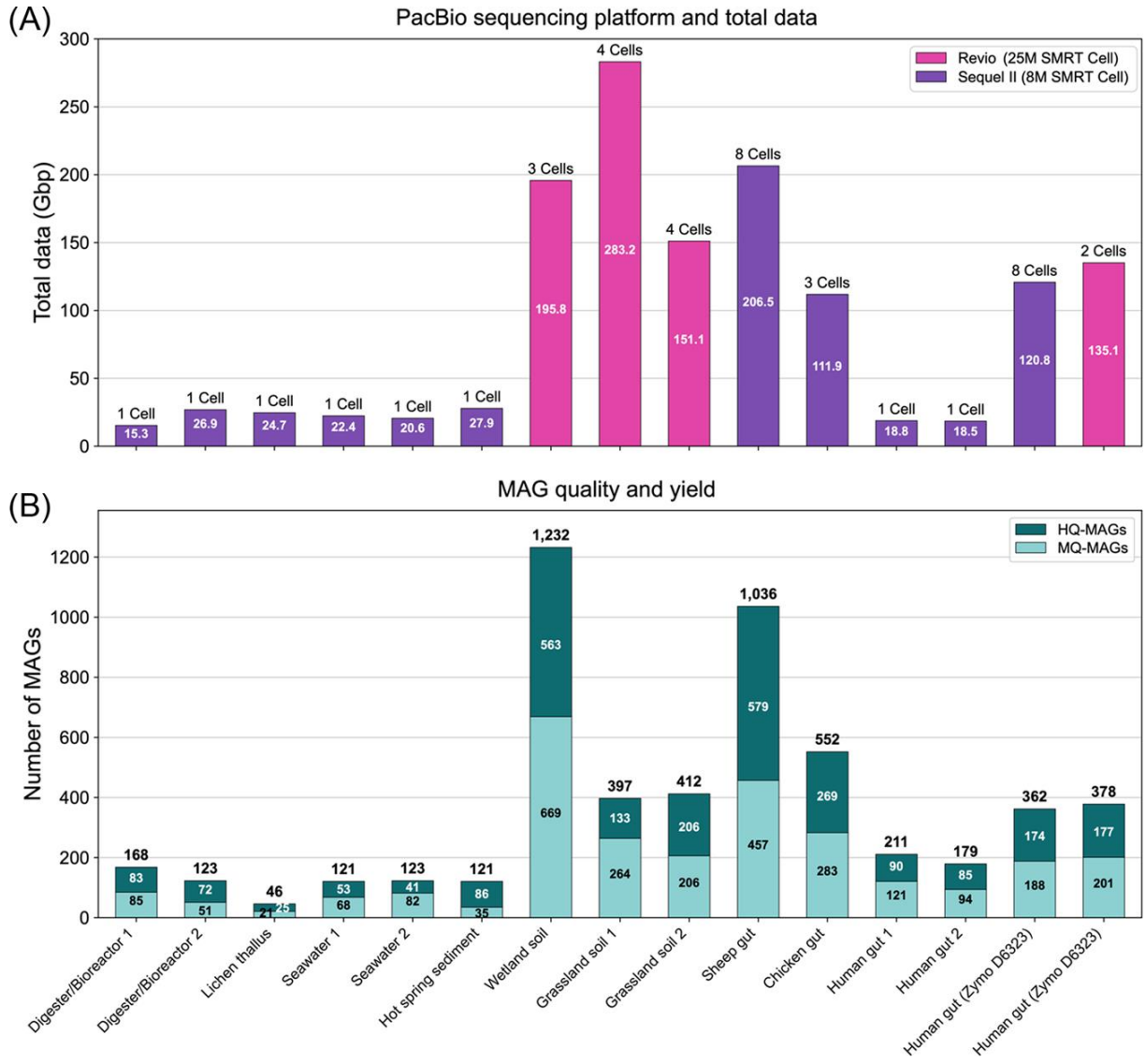


FIGURE 8.7 Sequencing data and number of MAGs across datasets. The total data for each sample are shown in panel (A), with Gbp labeled inside the bars and the number of SMRT cells used to generate the data labeled above the bars. Bars are colored by sequencing system used (Sequel II system: purple, Revio system: magenta). The total number of MAGs across samples is shown in panel (B). HQ-MAGs require ≥90% single-copy genes (SCGs) completeness and ≤5% contamination, whereas MQ-MAGs fall below the HQ thresholds but display ≥50% SCG completeness and ≤10% contamination. Total MAG

numbers are shown above (HQ-MAGs + MQ-MAGs), with colors showing MAG counts exclusive to each category.

Follow for extended description

The total number of MAGs recovered across all samples is shown in **Fig. 8.7** and reported in **Table 8.1** (along with the total data per sample). MAG yields vary across sample types and total data levels, but our Page 178 results can be used to set reasonable expectations for various types of experiments. For example there are eight datasets composed of a relatively low amount of total data (15–28 Gbp; **Fig. 8.7A**. These eight datasets were generated using one SMRT cell 8 M on the Sequel II system, but we note the same amount of data can be generated with a 5-plex on a Revio SMRT cell. For these eight datasets the total MAGs recovered ranges from 46 (lichen thallus) to 211 (human gut). The lichen thallus appears to be an outlier with relatively few MAGs, which is likely a reflection of the lower microbiome diversity. By contrast the remaining seven samples resulted in a minimum of 120 total MAGs (average = Page 179 149), including 41–90 HQ-MAGs each. These seven samples represent a diversity of microbiomes, including bioreactor, seawater, hot spring sediment, and human gut. Beyond these eight samples the other samples analyzed contained higher amounts of total data and were generated using multiple SMRT cells on the Sequel II or Revio systems (111–283 Gbp). These include the human (ZymoBIOMICS Fecal Reference D6323), chicken, and sheep gut samples, as well as the three different soil types. As expected these samples with higher total data resulted in higher numbers of total MAGs. Here the number of total MAGs ranged from 362 to 1232, which included 133–563 HQ-MAGs per sample (**Fig. 8.7B**). The wetland soil displayed the highest number of total MAGs (1232), followed by sheep gut (1036) and chicken gut (552). Beyond simply reporting the number of MAGs per sample we highlight two additional observations. First, for all 14 samples included in our analyses (both the low and high total data samples), the number of HQ-MAGs makes up an average of 48% of the

total MAGs (range: 33%–71%). These results indicate a substantial proportion of HQ-MAGs can be generated using HiFi sequencing, regardless of sample type. Second we successfully assembled many HQ-MAGs for the soil samples (range: 133–563; **Fig. 8.7B**). Soil has been notoriously difficult to assemble using short reads due to extremely high numbers of species occurring in relatively low abundances (Anthony et al., 2024). Our experiment demonstrates that it is possible to obtain many HQ-MAGs from soil samples, although it requires a higher depth of sequencing (e.g., multiple Revio SMRT cells). Together these results highlight how improved assembly methods, workflows, and HiFi reads can be used to obtain HQ-MAGs across a variety of host-associated and complex environmental microbiomes.

Our downsampling experiment sought to determine if there is a predictable relationship between total data and the number of MAGs recovered (including linear and nonlinear relationships). Our goal was to attempt to understand the relative proportion of genomes recovered per sample, and how many additional genomes could potentially be recovered with increased sequencing effort. We found a linear relationship for the three soil samples, whereas the pooled human gut (ZymoBIOMICS Fecal Reference D6323) displayed a nonlinear relationship resembling a saturation curve (**Fig. 8.8**). For the human gut sample the number of MAGs increases rapidly at first, but begins to level off with continued sequencing effort. Although additional MAGs are likely to be observed with additional sequencing, it seems that a

large proportion of genomes are already captured in the full dataset. By contrast, there is a linear relationship between total data and the number of MAGs recovered for each of the soil samples. These results suggest that a substantial proportion of the total genomes in the soil samples are not captured in the full datasets. In these soil samples a substantial amount of sequencing would likely need to occur before observing a saturation curve. We note that the highest species diversity was estimated for the grassland soil 1, followed by grassland soil 2, then wetland soil (based on 16 S amplicon analysis; not shown), and these diversity patterns also appear

to be reflected in the different slopes of the regression lines. More broadly the differences between the human gut and soil samples are driven by species/strain diversity and relative abundance patterns, with the human gut being significantly less complex than the soil. However the human gut sample included in this analysis is a pooled gut microbiome (e.g., containing multiple individuals), which is more complex than a typical gut microbiome from a single individual. We would therefore expect a typical gut microbiome to reach saturation with less total data than seen here for the pooled gut sample.

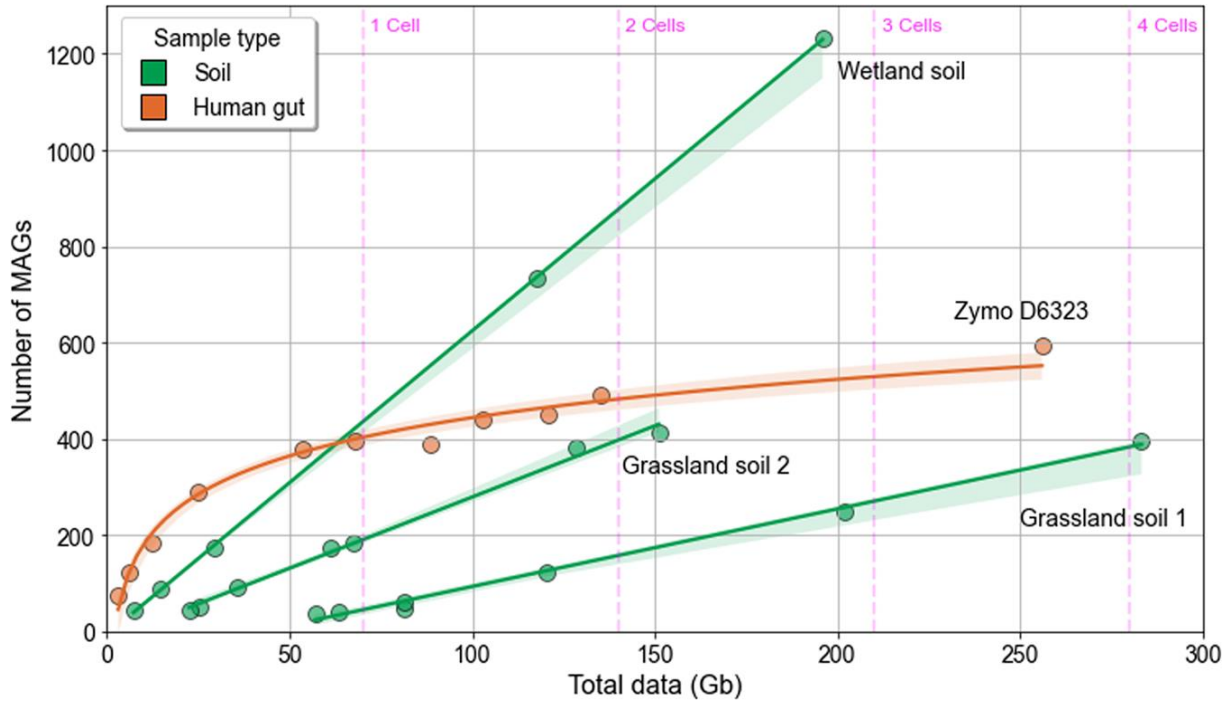


FIGURE 8.8 Effects of total data on MAG recovery. Number of total MAGs recovered across the downsampled datasets (soil = green color; pooled human gut = orange color). Dataset names correspond to those listed in **Table 8.1**. The expected total data for different numbers of Revio SMRT cells are shown by the magenta dotted lines.

Follow for extended description

When conducting an experiment focused on metagenome assembly it is not always clear how much sequencing effort is required to capture Page 181 most of the genomes available. Downsampling experiments, like the one conducted here, can provide valuable information on this topic. For example using one Revio SMRT cell would produce very few total MAGs for the grassland soil samples (**Fig. 8.8**). However this is not a failure of the assembly and binning methods themselves, but rather a lack of sufficient sequencing depth required to assemble the contigs. For any project we therefore recommend performing downsampling experiments based on the initial level of sequencing effort. The results can determine if a linear or nonlinear (e.g., saturation curve) relationship exists between total data and MAG yields, which can approximate the proportion of genomes assembled and help determine if additional sequencing is desired.

Advances in HiFi sequencing platforms are improving accessibility and costs for metagenome assembly, relative to prior projects completed with older platforms. For example the volume of data generated by the Revio system has enabled some of the deep-sequencing projects highlighted here. Importantly the sizes of these datasets (150–280 Gbp) were previously considered impractical and cost-prohibitive using the Sequel II system. For example the 150 Gbp dataset (grassland soil 2) would have required seven Sequel II SMRT cells 8 M, 9 days of sequencing time, and a consumable cost of ~$8000 USD. However, as of December 2024, a 150 Gbp dataset can now be generated using two SMRT cells on the Revio system (~$2000 USD consumable cost) or three SMRT cells on the Vega system (~$3300 USD consumable cost). The sequencing time is also dramatically reduced in this example—from 9 days on the Sequel II system to 24 hours on the Revio system (four SMRT cells can be run simultaneously) or 3 days on the Vega system (24 hour run times per SMRT cell, run sequentially). As a more extreme example the largest dataset included here was 283 Gbp (grassland soil 1), which would have required 14 Sequel II SMRT cells 8 M, 2.5 weeks of sequencing time, and ~$15,000 USD. This 283 Gbp deep-sequencing dataset

can now be generated using three Revio SMRT cells with SPRQ chemistry (~$3000 USD) in 24 hours or by using five Vega SMRT cells (~$5500 USD) in 5 days. Finally the Revio and Vega systems enable multiplexing of less complex sample types, further reducing costs for studies focused on metagenome assembly. Depending on the library prep and adapter kits used the cost to produce a barcoded library ranges from $39–$82. Success of multiplexing for metagenome assembly depends on the species diversity of the samples, including the total number of species and their relative Page 182 abundance profiles. Depending on the microbiomes it is possible to multiplex at a high level (e.g. 4–48 plex on the Revio system: artificial communities, fermentation, mock communities, etc.), whereas human gut microbiomes may perform well with fewer multiplexed samples (e.g. 2–16 plex on the Revio system). Before attempting multiplexing for metagenome assembly it is recommended to identify published results for a given microbiome sample (total data, number of MAGs) or perform a sequencing run with a single sample followed by a downsampling experiment (as described above). Regarding current costs across technologies Minich et al. (2024) sequenced the same set of human gut microbiomes with PacBio HiFi, ONT, and Illumina. They compared the total costs of sequencing, and the number of complete MAGs obtained per sample. They found that PacBio HiFi sequencing on the Revio system produced the greatest number of complete MAGs (2.4x more than ONT, 72x more than Illumina) and at the lowest cost. With excellent performance and cost parity the authors proposed HiFi sequencing will continue to gain traction for metagenome assembly.

Concluding remarks

Long-read sequencing has been transformative for targeted amplicon and metagenomic sequencing experiments (Agustinho et al., 2024). In particular the high accuracy of PacBio HiFi reads has facilitated the development of new methods and approaches that are driving the field forward. In this chapter we focused on the impact of HiFi sequencing on targeted amplicon sequencing

(16 S, ITS, etc.), taxonomic profiling of shotgun metagenomic reads, and metagenome assembly. For targeted amplicon sequencing it is clear that full-length 16 S sequencing outperforms partial 16 S sequencing and that high accuracy is required for high-precision taxonomic assignment (Johnson et al., 2019; Lefoulon et al., 2021; Notario et al., 2023; Buetas et al., 2024; Chuang et al., 2024). The same arguments apply to other amplicons, including fungal ITS (Tedersoo et al., 2018; Tedersoo et al., 2021; Latz et al., 2022; Runnel et al., 2022; Overgaard et al., 2024). Beyond standard FL 16 S or ITS sequencing, the PacBio Kinnex kit for 16 S enables significant throughput increases, with up to 12-fold improvements in yield. This approach allows higher levels of multiplexing, significantly reducing the cost per sample while also producing more reads per sample. For the taxonomic classification and profiling of metagenomic reads HiFi reads achieve substantially higher precision Page 183 relative to short reads (Portik et al., 2022). Long-read classifiers or profilers which leverage the large amount of information per HiFi read (typically spanning multiple genes and intergenic regions) produce highly accurate results, minimizing false positives while detecting nearly all species present. Consequently the field recognizes the power of long reads for this application, and there has been acceleration in the development of new taxonomic methods. Finally HiFi sequencing has revolutionized metagenome assembly, unlocking new opportunities to recover high-quality genomes from uncultured species in complex communities. New HiFi assembly methods and long-read specific workflows have pushed the limits of metagenome assembly, and large numbers of high-quality MAGs can now be recovered across a variety of host-associated and environmental microbiomes (Feng et al., 2022; Agustinho et al., 2024; Benoit et al., 2024; Portik et al., 2024). The availability of high-throughput HiFi sequencing is expected to accelerate biodiversity discovery, the characterization of genomes from uncultured microbes, and the continued development of new assembly and binning methods.

References

Acheampong et al., 2024 Acheampong DA, Jenjaroenpun P, Wongsurawat T, et al. CAIM: Coverage-based analysis for identification of microbiome. *bioRxiv* 2024; **https://doi.org/10.1101/2024.04.25.591018**.

Agustinho et al., 2024 Agustinho DP, Fu Y, Menon VK, Metcalf GA, Treangen TJ, Sedlazeck FJ. Unveiling microbial diversity: Harnessing long-read sequencing technology. *United States Nature Methods*. 2024;21(6):954–966 **https://doi.org/10.1038/s41592-024-02262-1**.

Ahmad et al., 2023 Ahmad N, Ritz M, Calchera A, et al. Biosynthetic potential of Hypogymnia holobionts: Insights into secondary metabolite pathways. *Journal of Fungi*. 2023;9546 **https://doi.org/10.3390/jof9050546**.

Al'Khafaji et al., 2024 Al'Khafaji AM, Smith JT, Garimella KV, et al. High-throughput RNA isoform sequencing using programmed cDNA concatenation. *United States Nature Biotechnology*. 2024;42(4):582–586 **https://doi.org/10.1038/s41587-023-01815-7**.

Almeida et al., 2019 Almeida A, Mitchell AL, Boland M, et al. A new genomic blueprint of the human gut microbiota. *Nature Publishing Group, United Kingdom Nature*. 2019;568(7753):499–504 **https://doi.org/10.1038/s41586-019-0965-1**.

Almeida et al., 2021 Almeida A, Nayfach S, Boland M, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *United Kingdom Nature Biotechnology*. 2021;39(1):105–114 **https://doi.org/10.1038/s41587-020-0603-3**.

Anthony et al., 2024 Anthony WE, Allison SD, Broderick CM, et al. From soil to sequence: Filling the critical gap in genome-resolved metagenomics is essential to the future of soil microbial ecology. *United States Environmental Microbiome*. 2024;19 **https://doi.org/10.1186/s40793-024-00599-w**.

Baldrian, 2019 Baldrian P. The known and the unknown in soil microbial ecology. *FEMS Microbiology Ecology*. 2019;95 **https://doi.org/10.1093/femsec/fiz005**.

Banerjee and van der Heijden, 2023 Banerjee S, van der Heijden MGA. Soil microbiomes and one health. *United States Nature Reviews Microbiology*. 2023;21(1):6–20 **https://doi.org/10.1038/s41579-022-00779-w**.

Benoit et al., 2024 Benoit G, Raguideau S, James R, Phillippy AM, Chikhi R, Quince C. High-quality metagenome assembly from long accurate reads with metaMDBG. *United Kingdom Nature Biotechnology*. 2024;42(9):1378–1383 **https://doi.org/10.1038/s41587-023-01983-6**.

Benoit et al., 2024 Benoit G, Raguideau S, James R, Phillippy AM, Chikhi R, Quince C. High-quality metagenome assembly from long accurate reads with metaMDBG. *United Kingdom Nature Biotechnology* 2024; **https://doi.org/10.1038/s41587-023-01983-6**.

Bickhart et al., 2022 Bickhart DM, Kolmogorov M, Tseng E, et al. Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. *United States Nature Biotechnology*. 2022;40(5):711–719 **https://doi.org/10.1038/s41587-021-01130-z**.

Blin et al., 2023 Blin K, Shaw S, Augustijn HE, et al. AntiSMASH 7.0: New and improved predictions for detection, regulation, chemical structures and visualisation. *Denmark Nucleic Acids Research*. 2023;51(1):W46 **https://doi.org/10.1093/nar/gkad344**.

Breitwieser et al., 2018 Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. *United States Briefings in Bioinformatics*. 2018;20(4):1125–1139 **https://doi.org/10.1093/bib/bbx120**.

Brown and Irber, 2016 Brown CT, Irber L. Sourmash: A library for MinHash sketching of DNA. *The Journal of Open Source Software*. 2016;1(5):27 **https://doi.org/10.21105/joss.00027**.

Buetas et al., 2024 Buetas E, Jordán-López M, López-Roldán A, et al. Full-length 16S rRNA gene sequencing by PacBio improves taxonomic resolution in human microbiome samples. *BMC Genomics*. 2024;25 **https://doi.org/10.1186/s12864-024-10213-5**.

Bui and Wei, 2020 Bui VK, Wei C. CDKAM: A taxonomic classification tool using discriminative k-mers and approximate matching strategies. *China BMC Bioinformatics*. 2020;21 **https://doi.org/10.1186/s12859-020-03777-y**.

Burton et al., 2014 Burton JN, Liachko I, Dunham MJ, Shendure J. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *United States G3: Genes, Genomes, Genetics*. 2014;4(7):1339–1346 **https://doi.org/10.1534/g3.114.011825**.

Chaumeil et al., 2020 Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. *Australia Bioinformatics*. 2020;36(6):1925–1927 **https://doi.org/10.1093/bioinformatics/btz848**.

Chaumeil et al., 2022 Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk v2: Memory friendly classification with the genome taxonomy database. *Australia Bioinformatics*. 2022;38(23):5315–5316 **https://doi.org/10.1093/bioinformatics/btac672**.

Chen et al., 2020 Chen LX, Anantharaman K, Shaiber A, Eren AM, Banfield JF. Accurate and complete genomes from metagenomes. *United States Genome Research*. 2020;30(3):315–333 **https://doi.org/10.1101/gr.258640.119**.

Chen et al., 2024 Chen X, Yin X, Shi X, Yan W, Yang Y, Liu L. Melon: metagenomic long-read-based taxonomic identification and quantification using marker genes. *Genome Biology*. 2024;25226 **https://doi.org/10.1186/s13059-024-03363-y**.

Chklovski et al., 2023 Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. CheckM2: A rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Australia Nature Methods*. 2023;20(8):1203–1212 **https://doi.org/10.1038/s41592-023-01940-w**.

Chuang et al., 2024 Chuang HH, Huang CG, Chou SH, Li HY, Lee CC, Lee LA. Comparative analysis of gut microbiota in children with obstructive sleep apnea: Assessing the efficacy of 16S rRNA gene sequencing in metabolic function prediction based on weight status. *Taiwan Frontiers in Endocrinology*. 2024;15 **https://doi.org/10.3389/fendo.2024.1344152**.

da Silva et al., 2024 da Silva RP, Suphavilai C, Nagarajan N. MetageNN: A memory-efficient neural network taxonomic classifier robust to sequencing errors and missing genomes. *BMC Bioinformatics*. 2024;25 **https://doi.org/10.1186/s12859-024-05760-3**.

Demaere and Darling, 2019 Demaere MZ, Darling AE. Bin3C: Exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes. *Australia Genome Biology*. 2019;20 **https://doi.org/10.1186/s13059-019-1643-1**.

Dilthey et al., 2019 Dilthey AT, Jain C, Koren S, Phillippy AM. Strain-level metagenomic assignment and compositional estimation for

long reads with MetaMaps. *Nature Communications*. 2019;10 **https://doi.org/10.1038/s41467-019-10934-2**.

Duvallet et al., 2017 Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature Communications*. 2017;8 **https://doi.org/10.1038/s41467-017-01973-8**.

Eisenhofer et al., 2024 Eisenhofer R, Nesme J, Santos-Bay L, et al. A comparison of short-read, HiFi long-read, and hybrid strategies for genome-resolved metagenomics. *Microbiology Spectrum*. 2024;12 **https://doi.org/10.1128/spectrum.03590-23**.

Fan et al., 2021 Fan J, Huang S, Chorlton SD. BugSeq: A highly accurate cloud platform for long-read metagenomic analyses. *BMC Bioinformatics*. 2021;22 **https://doi.org/10.1186/s12859-021-04089-5**.

Feng et al., 2022 Feng X, Cheng H, Portik D, Li H. Metagenome assembly of high-fidelity long reads with hifiasm-meta. *Nature Methods*. 2022;19(6):671–674 **https://doi.org/10.1038/s41592-022-01478-3**.

Forster et al., 2019 Forster SC, Kumar N, Anonye BO, et al. A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nature Biotechnology*. 2019;37(2):186–192 **https://doi.org/10.1038/s41587-018-0009-7**.

Gao et al., 2024 Gao J, Wang Z, Deng W, et al. Improved resolution of microbial diversity in deep-sea surface sediments using PacBio long-read 16S rRNA gene sequencing. *China mSphere*. 2024;9 **https://doi.org/10.1128/msphere.00770-24**.

Gehrig et al., 2022 Gehrig JL, Portik DM, Driscoll MD, et al. Finding the right fit: Evaluation of short-read and long-read sequencing approaches to maximize the utility of clinical microbiome data. *United States Microbial Genomics*. 2022;8 **https://doi.org/10.1099/mgen.0.000794**.

Gilbert et al., 2018 Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. *United States Nature Medicine*. 2018;24(4):392–400 **https://doi.org/10.1038/nm.4517**.

Hirsch et al., 2024 Hirsch P, Molano LAG, Engel A, et al. Mibianto: Ultra-efficient online microbiome analysis through k-mer based metagenomics. *Germany Nucleic Acids Research*. 2024;52(1):W407 **https://doi.org/10.1093/nar/gkae364**.

Huson et al., 2018 Huson DH, Albrecht B, Bağci C, et al. MEGAN-LR: New algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Germany Biology Direct*. 2018;13 **https://doi.org/10.1186/s13062-018-0208-7**.

Hyatt et al., 2010 Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. 2010;11 **http://doi.org/10.1186/1471-2105-11-119**.

Irber et al., 2022 Irber L, Brooks PT, Reiter T, et al. Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers. *United States bioRxiv* 2022; **https://doi.org/10.1101/2022.01.11.475838**.

Jiang et al., 2023 Jiang F, Li Q, Wang S, et al. Recovery of metagenome-assembled microbial genomes from a full-scale biogas plant of food waste by pacific biosciences high-fidelity sequencing. *Frontiers in Microbiology*. 2023;13 **https://doi.org/10.3389/fmicb.2022.1095497**.

Johnson et al., 2019 Johnson JS, Spakowicz DJ, Hong BY, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *United States Nature Communications*. 2019;10 **https://doi.org/10.1038/s41467-019-13036-1**.

Kang et al., 2019 Kang DD, Li F, Kirton E, et al. MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *Peer Journal*. 2019;7(7):e7359 **https://doi.org/10.7717/peerj.7359**.

Kato et al., 2022 Kato S, Masuda S, Shibata A, Shirasu K, Ohkuma M. Insights into ecological roles of uncultivated bacteria in Katase hot spring sediment from long-read metagenomics. *Frontiers in Microbiology*. 2022;13 **https://doi.org/10.3389/fmicb.2022.1045931**.

Kim et al., 2022 Kim CY, Ma J, Lee I. HiFi metagenomic sequencing enables assembly of accurate and complete genomes from human gut microbiota. *South Korea Nature Communications*. 2022;13 **https://doi.org/10.1038/s41467-022-34149-0**.

Kim and Steinegger, 2024 Kim J, Steinegger M. Metabuli: Sensitive and specific metagenomic classification via joint analysis of amino acid and DNA. *Nature Methods*. 2024;21(6):971–973 **https://doi.org/10.1038/s41592-024-02273-y**.

Kolmogorov et al., 2020 Kolmogorov M, Bickhart DM, Behsaz B, et al. metaFlye: Scalable long-read metagenome assembly using repeat graphs. *United States Nature Methods*. 2020;17(11):1103–1110 **https://doi.org/10.1038/s41592-020-00971-x**.

Kutuzova et al., 2024 Kutuzova S, Piera P, Nielsen KN, et al. Binning meets taxonomy: TaxVAMB improves metagenome binning using bi-modal variational autoencoder. *Denmark bioRxiv* 2024; **https://doi.org/10.1101/2024.10.25.620172**.

Kyrpides et al., 2017 Kyrpides NC, Stepanauskas R, Harmon-Smith M, et al.... , Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *United States Nature Biotechnology*. 2017;35(8):725–731 **https://doi.org/10.1038/nbt.3893**.

Lagier et al., 2018 Lagier JC, Dubourg G, Million M, et al. Culturing the human microbiota and culturomics. *France Nature Reviews Microbiology*. 2018;16(9):540–550 **https://doi.org/10.1038/s41579-018-0041-0**.

Lamurias et al., 2022 Lamurias A, Sereika M, Albertsen M, Hose K, Nielsen TD. Metagenomic binning with assembly graph embeddings. *Denmark Bioinformatics*. 2022;38(19):4481–4487 **https://doi.org/10.1093/bioinformatics/btac557**.

Latz et al., 2022 Latz MAC, Grujcic V, Brugel S, et al. Short- and long-read metabarcoding of the eukaryotic rRNA operon: Evaluation of primers and comparison to shotgun metagenomics sequencing. *Sweden Molecular Ecology Resources*. 2022;22(6):2304–2318 **https://doi.org/10.1111/1755-0998.13623**.

Lefoulon et al., 2021 Lefoulon E, Truchon A, Clark T, Long C, Frey D, Slatko BE. Greenhead (tabanus nigrovittatus) wolbachia and its microbiome: A preliminary study. *United States Microbiology Spectrum*. 2021;9 **https://doi.org/10.1128/Spectrum.00517-21**.

Lloyd-Price et al., 2019 Lloyd-Price J, Arze C, Ananthakrishnan AN, et al.... Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *United States Nature*. 2019;569(7758):655–662 **https://doi.org/10.1038/s41586-019-1237-9**.

Lynch and Pedersen, 2016 Lynch SV, Pedersen O. The human intestinal microbiome in health and disease. *New England Journal of Medicine*. 2016;375(24):2369–2379 **https://doi.org/10.1056/NEJMra1600266**.

Mainguy et al., 2024 Mainguy J, Vienne M, Fourquet J, et al. metagWGS: A comprehensive workflow to analyze metagenomic

data using Illumina or PacBio HiFi reads. *France bioRxiv* 2024; **https://doi.org/10.1101/2024.09.13.612854**.

Mari et al., 2024 Mari J, Križanovi K, Riondet S, Nagarajan N, Šiki M. Comparative analysis of metagenomic classifiers for long-read sequencing datasets. *Croatia BMC Bioinformatics*. 2024;25 **https://doi.org/10.1186/s12859-024-05634-8**.

Masuda et al., 2024 Masuda S, Gan P, Kiguchi Y, et al. Uncovering microbiomes of the rice phyllosphere using long-read metagenomic sequencing. *Communications Biology*. 2024;7 **https://doi.org/10.1038/s42003-024-05998-w**.

Meslier et al., 2022 Meslier V, Quinquis B, Da Silva K, et al. Benchmarking second and third-generation sequencing platforms for microbial metagenomics. *Scientific Data*. 2022;9 **https://doi.org/10.1038/s41597-022-01762-z**.

Meyer et al., 2022 Meyer F, Fritz A, Deng ZL, et al.... Critical Assessment of metagenome interpretation: The second round of challenges. *Germany Nature Methods*. 2022;19(4):429–440 **https://doi.org/10.1038/s41592-022-01431-4**.

Minich et al., 2024 Minich JJ, Allsing N, Din MO, et al. Culture-independent meta-pangenomics enabled by long-read metagenomics reveals novel associations with pediatric undernutrition. *United States SSRN* 2024; **https://doi.org/10.2139/ssrn.4960739**.

Mirdita et al., 2021 Mirdita M, Steinegger M, Breitwieser F, Söding J, Levy Karin E. Fast and sensitive taxonomic assignment to metagenomic contigs. *Germany Bioinformatics*. 2021;37(18):3029–3031 **https://doi.org/10.1093/bioinformatics/btab184**.

Nayfach et al., 2019 Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC. New insights from uncultivated genomes of the global human gut microbiome. *United States Nature*.

2019;568(7753):505–510 **https://doi.org/10.1038/s41586-019-1058-x**.

Notario et al., 2023 Notario E, Visci G, Fosso B, et al. Amplicon-based microbiome profiling: From second- to third-generation sequencing for higher taxonomic resolution. *Genes*. 2023;14(8):1567 **https://doi.org/10.3390/genes14081567**.

Olm et al., 2017 Olm MR, Brown CT, Brooks B, Banfield JF. DRep: A tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *United States ISME Journal*. 2017;11(12):2864–2868 **https://doi.org/10.1038/ismej.2017.126**.

Orellana et al., 2023 Orellana LH, Krüger K, Sidhu C, Amann R. Comparing genomes recovered from time-series metagenomes using long- and short-read sequencing technologies. *Microbiome*. 2023;11 **https://doi.org/10.1186/s40168-023-01557-3**.

Overgaard et al., 2024 Overgaard CK, Jamy M, Radutoiu S, Burki F, Dueholm MKD. Benchmarking long-read sequencing strategies for obtaining ASV-resolved rRNA operons from environmental microeukaryotes. *Molecular Ecology Resources*. 2024;24 **https://doi.org/10.1111/1755-0998.13991**.

Pan et al., 2023 Pan S, Zhao XM, Coelho LP. SemiBin2: Self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing. *Bioinformatics (Oxford, England)*. 2023;39:I21 **https://doi.org/10.1093/bioinformatics/btad209**.

Pasolli et al., 2019 Pasolli E, Asnicar F, Manara S, et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*. 2019;176(3):649 **https://doi.org/10.1016/j.cell.2019.01.001**.

Pierce et al., 2019 Pierce NT, Irber L, Reiter T, Brooks P, Brown CT. Large-scale sequence comparisons with sourmash.

F1000Research. 2019;8 **https://doi.org/10.12688/f1000research.19675.1**.

Portik et al., 2022 Portik DM, Brown CT, Pierce-Ward NT. Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets. *BMC Bioinformatics*. 2022;23 **https://doi.org/10.1186/s12859-022-05103-0**.

Portik et al., 2024 Portik DM, Feng X, Benoit G, et al. Highly accurate metagenome-assembled genomes from human gut microbiota using long-read assembly, binning, and consolidation methods. *bioRxiv* 2024; **https://doi.org/10.1101/2024.05.10.593587**.

Press et al., 2017 Press MO, Wiser AH, Kronenberg ZN, et al. Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions. *bioRxiv* 2017; **https://doi.org/10.1101/198713**.

Priest et al., 2021 Priest T, Orellana LH, Huettel B, Fuchs BM, Amann R. Microbial metagenome-assembled genomes of the Fram Strait from short and long read sequencing platforms. *Peer Journal*. 2021;9 **https://doi.org/10.7717/peerj.11721**.

Rho et al., 2010 Rho M, Tang H, Ye Y. FragGeneScan: Predicting genes in short and error-prone reads. *Nucleic Acids Research*. 2010;38(20):e191 **https://doi.org/10.1093/nar/gkq747**.

Runnel et al., 2022 Runnel K, Abarenkov K, Copoţ O, et al. DNA barcoding of fungal specimens using PacBio long-read high-throughput sequencing. *Molecular Ecology Resources*. 2022;22(8):2871–2879 **https://doi.org/10.1111/1755-0998.13663**.

Saak et al., 2023 Saak CC, Pierce EC, Dinh CB, et al. Longitudinal, multi-platform metagenomics yields a high-quality genomic catalog and guides an in vitro model for cheese communities. *United States mSystems*. 2023;8 **https://doi.org/10.1128/msystems.00701-22**.

Sapoval et al., 2024 Sapoval N, Liu Y, Curry KD, et al. Lightweight taxonomic profiling of long-read metagenomic datasets with Lemur and Magnet. *bioRxiv* 2024; **https://doi.org/10.1101/2024.06.01.596961**.

Schaerer et al., 2023 Schaerer L, Putman L, Bigcraft I, et al. Coexistence of specialist and generalist species within mixed plastic derivative-utilizing microbial communities. *Microbiome*. 2023;11 **https://doi.org/10.1186/s40168-023-01645-4**.

Sczyrba et al., 2017 Sczyrba A, Hofmann P, Belmann P, et al.... Critical Assessment of Metagenome Interpretation—A benchmark of metagenomics software. *Germany Nature Methods*. 2017;14(11):1063–1071 **https://doi.org/10.1038/nmeth.4458**.

Sereika et al., 2022 Sereika M, Kirkegaard RH, Karst SM, et al. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nature Methods*. 2022;19(7):823–826 **https://doi.org/10.1038/s41592-022-01539-7**.

Shaw and Yu, 2024 Shaw J, Yu YW. Rapid species-level metagenome profiling and containment estimation with sylph. *Nature Biotechnology* 2024; **https://doi.org/10.1038/s41587-024-02412-y**.

Sieber et al., 2018 Sieber CMK, Probst AJ, Sharrar A, et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology*. 2018;3(7):836–843 **https://doi.org/10.1038/s41564-018-0171-1**.

Tao et al., 2023 Tao Y, Xun F, Zhao C, et al. Improved assembly of metagenome-assembled genomes and viruses in Tibetan saline lake sediment by HiFi metagenomic sequencing. *Microbiology*

Spectrum. 2023;11 **https://doi.org/10.1128/spectrum.03328-22**.

Tedersoo et al., 2021 Tedersoo L, Albertsen M, Anslan S, Callahan B, Druzhinina IS. Perspectives and benefits of high-throughput long-read sequencing in microbial ecology. *Applied and Environmental Microbiology*. 2021;87 **https://doi.org/10.1128/aem.00626-21**.

Tedersoo et al., 2018 Tedersoo L, Tooming-Klunderud A, Anslan S. PacBio metabarcoding of fungi and other eukaryotes: Errors, biases and perspectives. *New Phytologist*. 2018;217(3):1370–1385 **https://doi.org/10.1111/nph.14776**.

Ulrich and Renard, 2024 Ulrich JU, Renard BY. Fast and space-efficient taxonomic classification of long reads with hierarchical interleaved XOR filters. *Genome Research*. 2024;34(6):914–924 **https://doi.org/10.1101/gr.278623.123**.

Vitorino and Bessa, 2018 Vitorino LC, Bessa LA. Microbial diversity: The gap between the estimated and the known. *Diversity*. 2018;10 **https://doi.org/10.3390/D10020046**.

Wang and Jia, 2016 Wang J, Jia H. Metagenome-wide association studies: Fine-mining the microbiome. *Nature Reviews Microbiology*. 2016;14(8):508–522 **https://doi.org/10.1038/nrmicro.2016.83**.

Wang et al., 2024 Wang YC, Mao Y, Fu HM, et al. New insights into functional divergence and adaptive evolution of uncultured bacteria in anammox community by complete genome-centric analysis. *Science of the Total Environment*. 2024;924 **https://doi.org/10.1016/j.scitotenv.2024.171530**.

Wenger et al., 2019 Wenger AM, Peluso P, Rowell WJ, et al.... Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature*

Biotechnology. 2019;37(10):1155–1162 **https://doi.org/10.1038/s41587-019-0217-9**.

Wickramarachchi et al., 2020 Wickramarachchi A, Mallawaarachchi V, Rajan V, Lin Y. MetaBCC-LR: Meta genomics binning by coverage and composition for long reads. *Bioinformatics (Oxford, England)*. 2020;36(Supplement_1):i3 **https://doi.org/10.1093/bioinformatics/btaa441**.

Wilkinson et al., 2021 Wilkinson JE, Franzosa EA, Everett C, et al.... A framework for microbiome science in public health. *United States Nature Medicine*. 2021;27(5):766–774 **https://doi.org/10.1038/s41591-021-01258-0**.

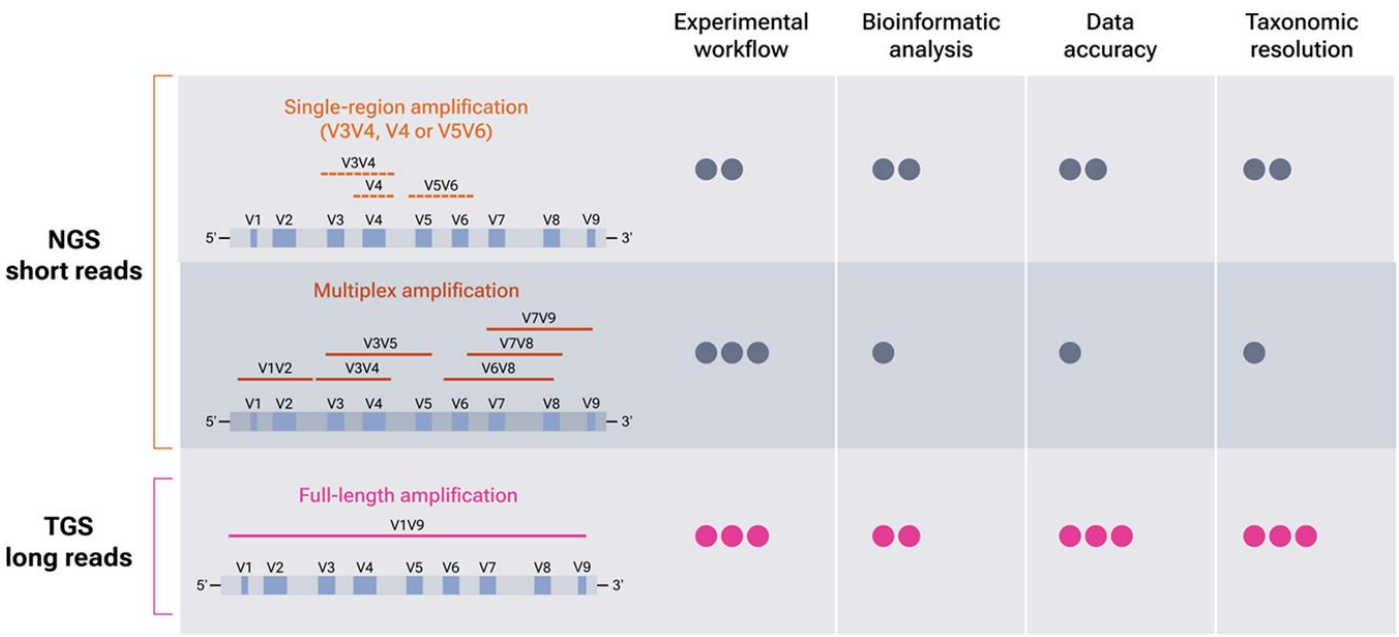
Yadav and Subramanian, 2024 Yadav A, Subramanian S. HiFiBGC: An ensemble approach for improved biosynthetic gene cluster detection in PacBio HiFi-read metagenomes. *BMC Genomics*. 2024;25 **https://doi.org/10.1186/s12864-024-10950-7**.

Zhang et al., 2024 Zhang W, Liu Y, Xu J, Chen E, Schonhuth A, Luo X. PanTax: Strain-level taxonomic classification of metagenomic data using pangenome graphs. *bioRxiv* 2024; **https://doi.org/10.1101/2024.11.15.623887**.

Zhang et al., 2023 Zhang Z, Yang C, Veldsman WP, Fang X, Zhang L. Benchmarking genome assembly methods on metagenomic sequencing data. *Briefings in Bioinformatics*. 2023;24 **https://doi.org/10.1093/bib/bbad087**.

Zou et al., 2019 Zou Y, Xue W, Luo G, et al.... Liang Xiao, 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nature Biotechnology*. 2019;37(2):179–185 **https://doi.org/10.1038/s41587-018-0008-8**.

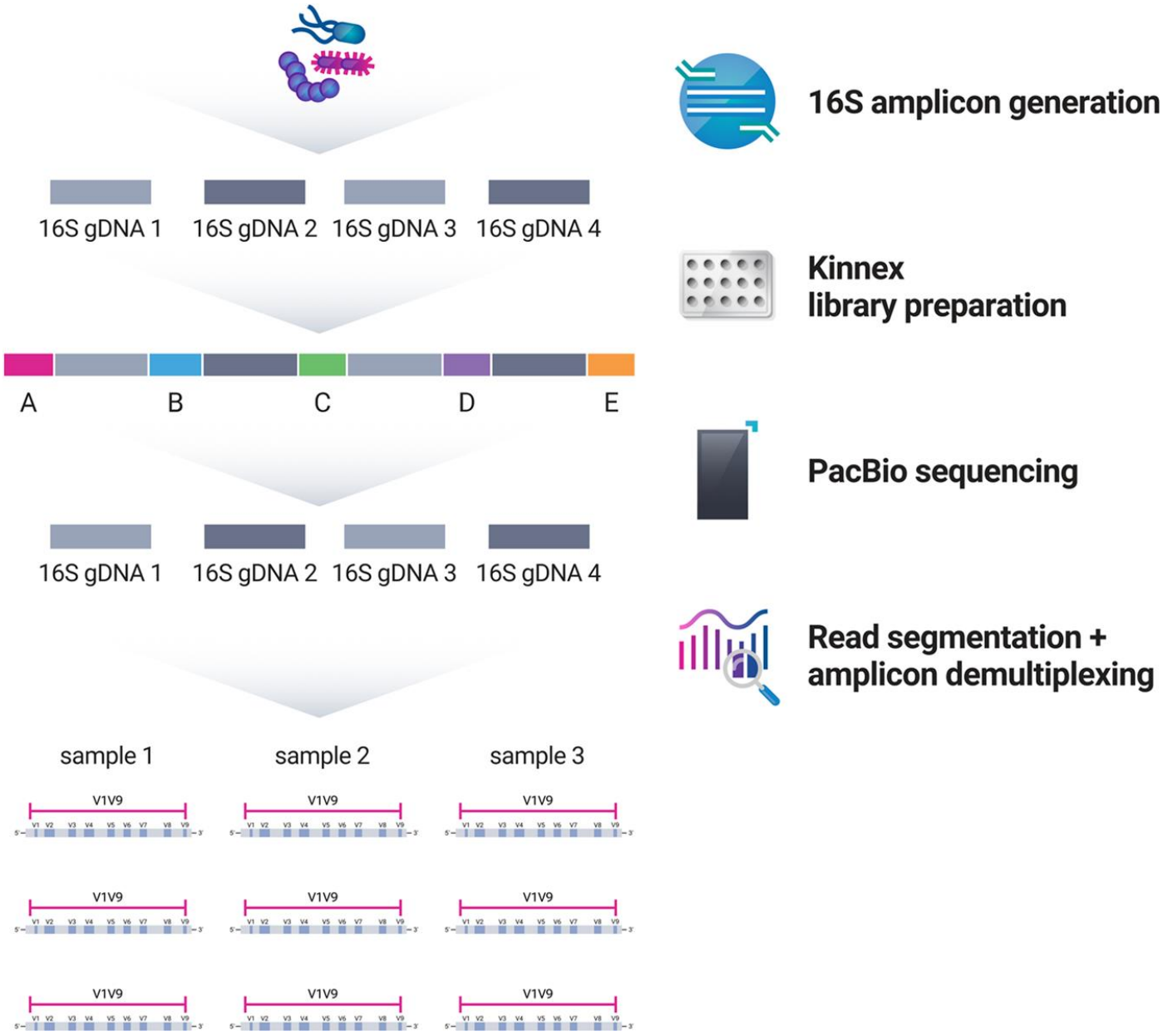
Description of Figure 8.1



The table consists of five columns: Amplification method, Experimental workflow, Bioinformatic analysis, Data accuracy and Taxonomic resolution. It reads as follows: Row 1. Amplification method: Single-region amplification (with V3V4, V4 or V5V6). Experimental workflow: two gray dots. Bioinformatic analysis: two gray dots. Data accuracy: two gray dots. Taxonomic resolution: two gray dots. Row 2. Amplification method: Multiplex amplification (with V1V2, V3V4, V3V5, V7V9, V7V8 or V6V8). Experimental workflow: three gray dots. Bioinformatic analysis: one gray dot. Data accuracy: one gray dot. Taxonomic resolution: one gray dot. Row 3. Amplification method: Full-length amplification (with V1V9). Experimental workflow: three pink dots. Bioinformatic analysis: three pink dots. Data accuracy: three pink dots. Taxonomic resolution: three pink dots.

Navigate back to Figure 8.1

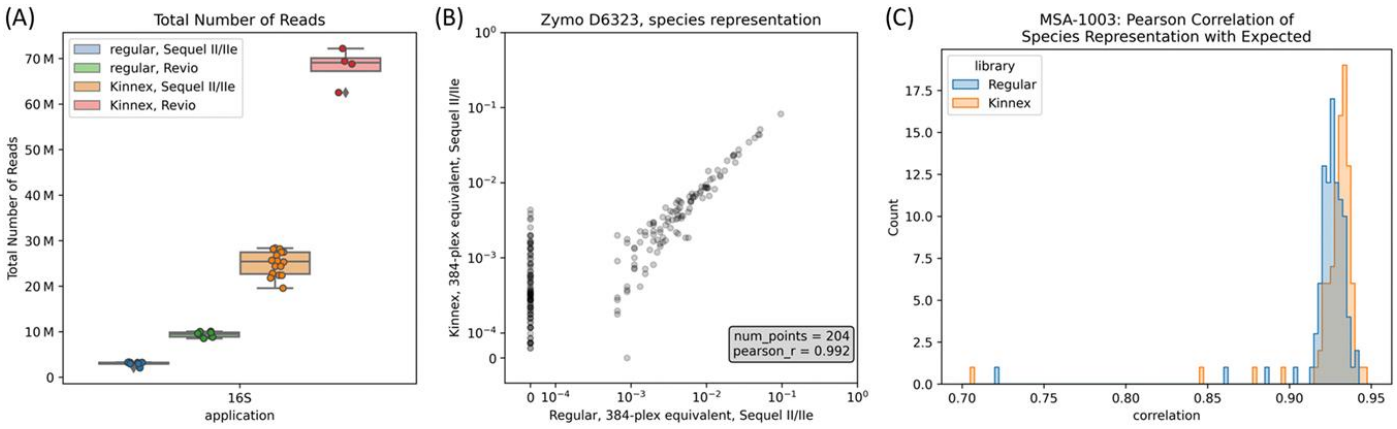
Description of Figure 8.2



The image shows a flowchart with the following steps: 16S amplicon generation, Kinnex library preparation, PacBio sequencing and read segmentation with amplicon demultiplexing. The process starts with 16S g D N A 1, 16S g D N A 2, 16S g D N A 3 and 16S g D N A 4. These are then processed through the steps mentioned. The final step shows sample 1, sample 2 and sample 3 with V1V9 regions.

Navigate back to Figure 8.2

Description of Figure 8.3

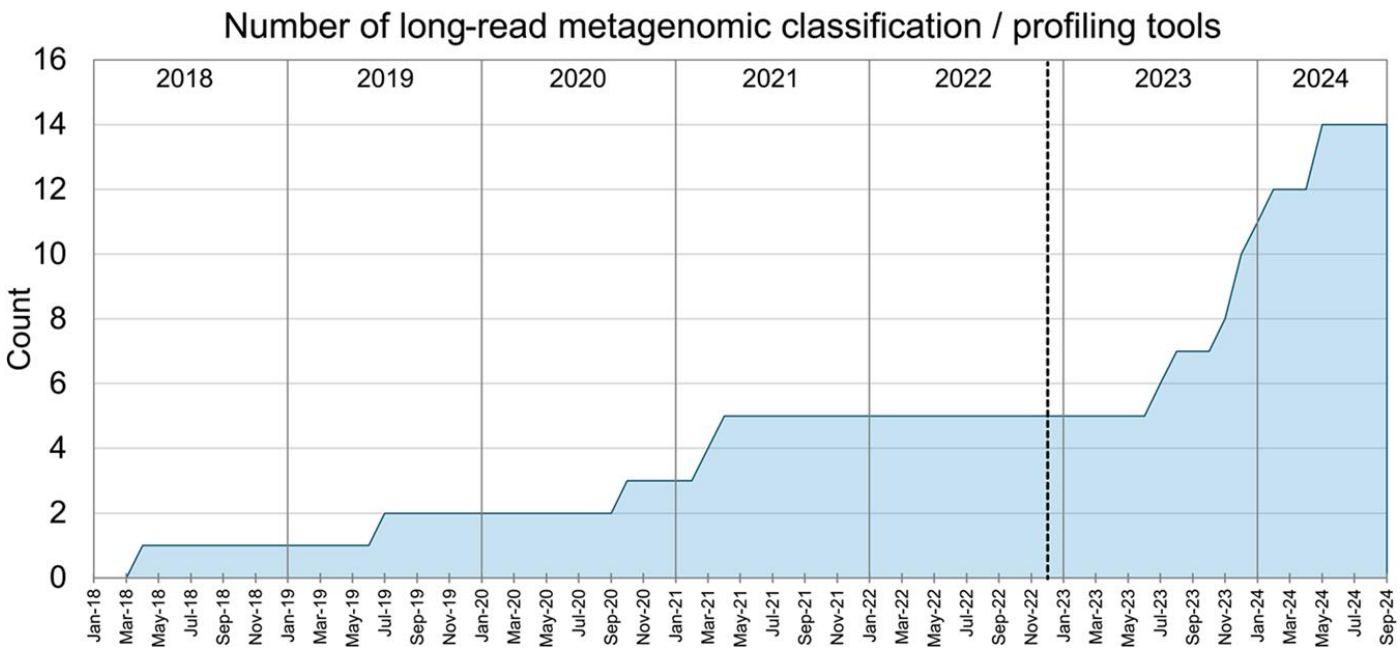


The image contains three graphs labeled A, B and C. Graph A: Total Number of Reads. The x-axis represents the application, showing 16S. The y-axis shows the Total Number of Reads, ranging from 0 to 70 M. Four box plots are displayed: 1. regular, Sequel 2 or 2e: approximately 2 to 3M reads 2. regular, Revio: approximately 8 to 10 M reads 3. Kinnex, Sequel 2 or 2e: approximately 20 to 30M reads 4. Kinnex, Revio: approximately 65 to 70M reads. Graph B: Zymo D6323, species representation. The x-axis represents Regular, 384-plex equivalent, Sequel 2 or 2e, ranging from 0 to 10 superscript negative 4 end superscript to 10 superscript 0 end superscript. The y-axis represents Kinnex, 384-plex equivalent, Sequel 2 or 2e, with the same range. The graph shows a scatter plot with a strong positive correlation. A text box in the lower right corner states: num underscore points equals 204, pearson underscore r equals 0.992. Graph C: MSA-1003: Pearson Correlation of Species Representation with Expected The x-axis shows the correlation from 0.70 to 0.95. The y-axis represents the Count ranging from 0.0 to 17.5. Two overlapping histograms are shown: 1. Regular library: peaks around 0.92 to 0.93 2. Kinnex library: peaks around 0.93 to 0.94 The graphs demonstrate a comparison between regular and Kinnex sequencing methods, showing differences in read counts and similarities in species representation. Various

smaller bars are present for 1 count at different correlation values for both regular and kinnex library.

Navigate back to Figure 8.3

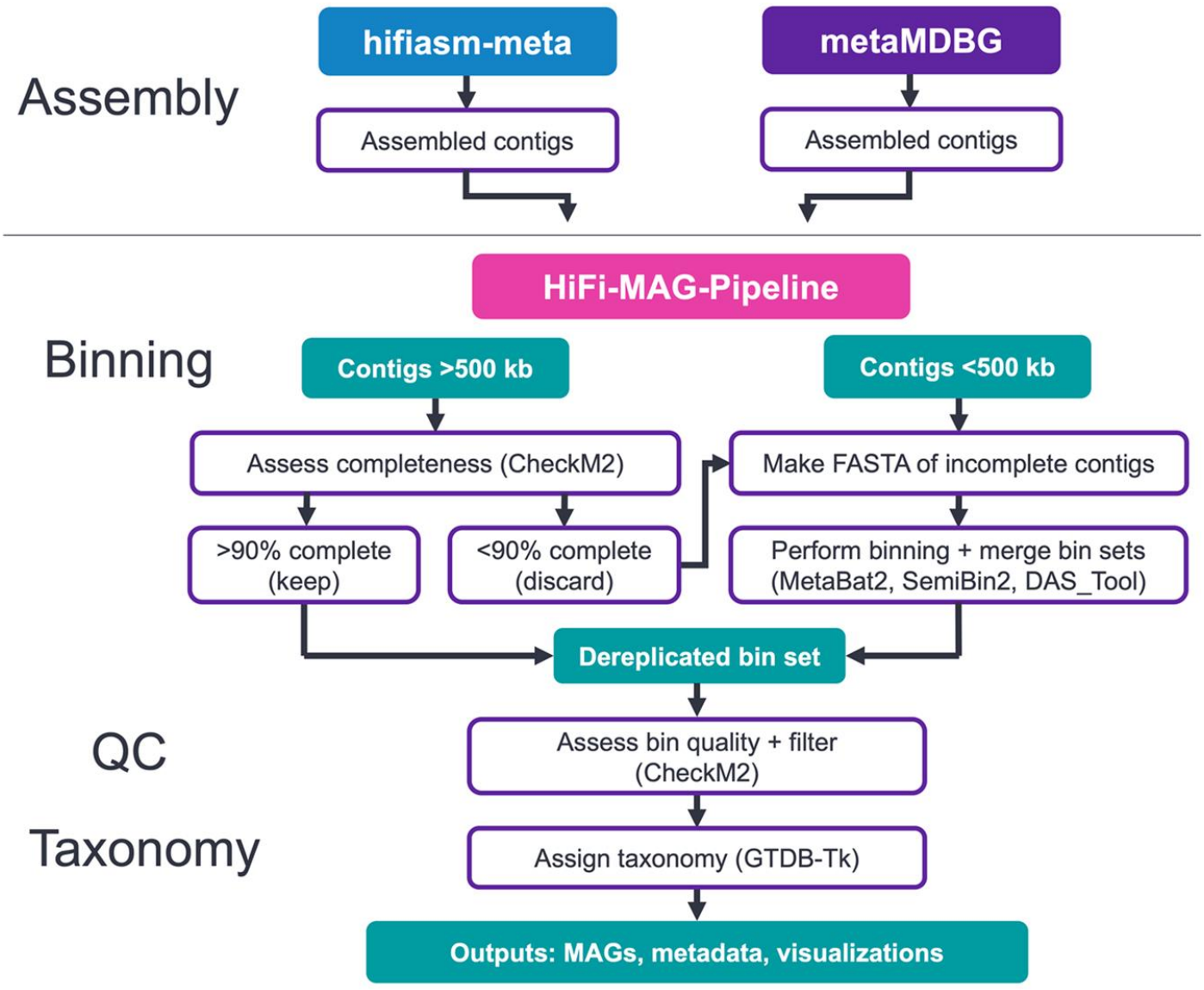
Description of Figure 8.4



The graph is titled 'Number of long-read metagenomic classification or profiling tools'. The x-axis represents time in months and years from January 2018 to September 2024. The y-axis shows the count of tools, ranging from 0 to 16 in increments of 2. The line graph starts at 0 and increases to 1 in March 2018 and remains flat until around May 2019, where it increases to 2. It stays at this level until September 2020, then rises to 3. In March 2021, it increases to 5 and remains steady until July 2023. From July 2023, there's a marked increase in the number of tools as follows: till September 2023 count is 7, count increases to 12 after January 2024 and remains same till may 2024 and finally increases to 14 after may 2024 and remains the same till the end at september 2024. A vertical dashed line separates the years 2022 and 2023, highlighting the significant increase in tools from 2023 onwards.

Navigate back to Figure 8.4

Description of Figure 8.6

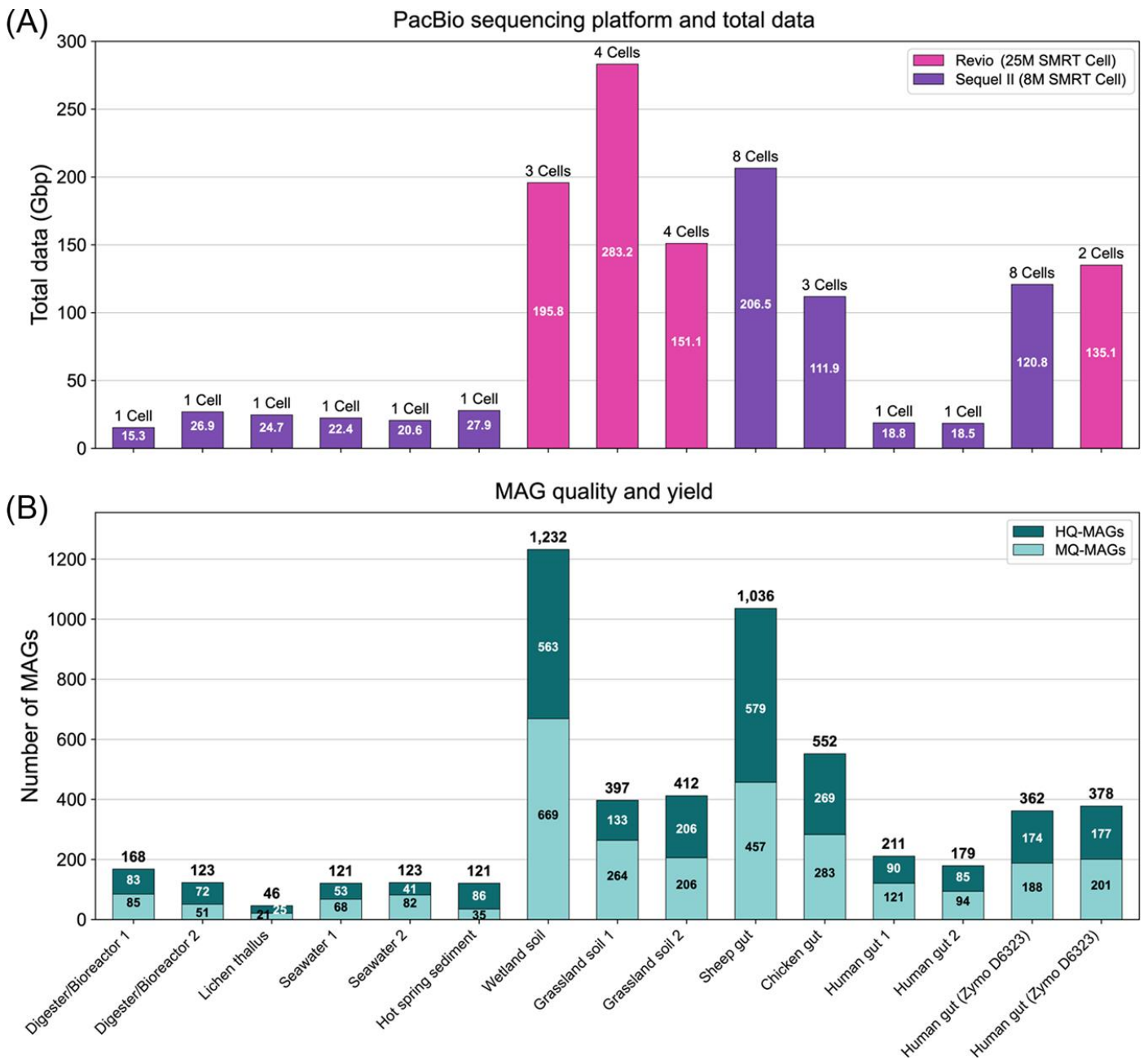


The process starts with Assembly, where 'hifiasm-meta' and 'metaM D B G' produce Assembled contigs. The Binning stage which is titled HiFi-MAG pipeline is where contigs are divided into those greater than 500 k b and those less than 500 k b. Contigs greater than 500 k b are assessed for completeness using Check M 2. If they are more than 90 percent complete, they are kept; if less than 90 percent complete, they are discarded. Contigs less than 500 k b are made into FASTA of incomplete contigs and then

undergo binning and merging using Meta Bat2, Semi Bin2 and DAS underscore Tool. The less than 90 percent complete discarded contigs are made into FASTA of incomplete contigs. Both paths lead to a Dereplicated bin set. In the Q C and Taxonomy stage, the bin quality is assessed and filtered using Check M 2 and taxonomy is assigned using G T D B-T k. The outputs are M A G s, metadata and visualizations.

Navigate back to Figure 8.6

Description of Figure 8.7



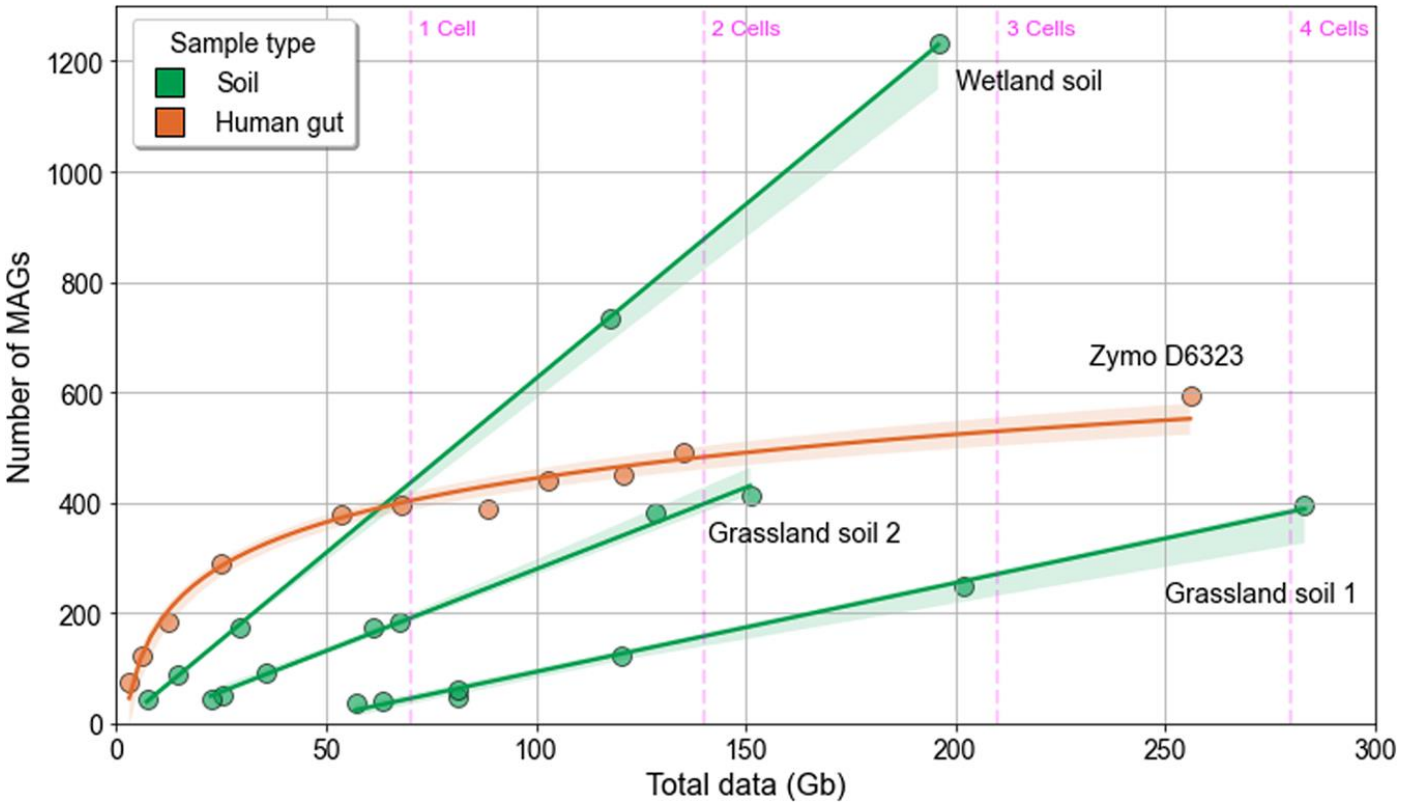
The image contains two bar graphs labeled A and B. Graph A is titled 'PacBio sequencing platform and total data'. The x-axis shows different conditions and the y-axis represents 'Total data (G b p)' ranging from 0 to 300. The graph compares two platforms: 1. Revio (25M S M R T Cell) 2. Sequel 2 (8M S M R T Cell). The detailed data points are as follows: For the PacBio sequencing

platform and total data: For the Revio (25M S M R T Cell) platform, with 1 cell, the total data is 26.9 G b p. For the Revio (25M S M R T Cell) platform, with 1 cell, the total data is 20.6 G b p. For the Revio (25M S M R T Cell) platform, with 3 cells, the total data is 195.8 G b p. For the Revio (25M S M R T Cell) platform, with 4 cells, the total data is 283.2 G b p. For the Revio (25M S M R T Cell) platform, with 4 cells, the total data is 151.1 G b p. For the Revio (25M S M R T Cell) platform, with 2 cells, the total data is 135.1 G b p. For the Sequel 2 (8M S M R T Cell) platform, with 1 cell, the total data is 15.3 G b p. For the Sequel 2 (8M S M R T Cell) platform, with 1 cell, the total data is 24.7 G b p. For the Sequel 2 (8M S M R T Cell) platform, with 1 cell, the total data is 22.4 G b p. For the Sequel 2 (8M S M R T Cell) platform, with 1 cell, the total data is 27.9 G b p. For the Sequel 2 (8M S M R T Cell) platform, with 8 cells, the total data is 206.5 G b p. For the Sequel 2 (8M S M R T Cell) platform, with 3 cells, the total data is 111.9 G b p. For the Sequel 2 (8M S M R T Cell) platform, with 1 cell, the total data is 18.8 G b p. For the Sequel 2 (8M S M R T Cell) platform, with 1 cell, the total data is 18.5 G b p. For the Sequel 2 (8M S M R T Cell) platform, with 8 cells, the total data is 120.8 G b p. Graph B is titled 'M A G quality and yield'. The x-axis shows different sample types and the y-axis represents 'Number of M A G s' ranging from 0 to 1200. The graph compares two types of M A G s: 1. H Q M A G s (High-Quality M A G s) 2. M Q M A G s (Medium-Quality M A G s) Data points for various samples for the M A G quality and yield are as follows: For the Digester or Bioreactor 1 sample, there are 83 H Q M A G s, 85 M Q M A G s and a total of 168 M A G s. For the Digester or Bioreactor 2 sample, there are 72 H Q M A G s, 51 M Q M A G s and a total of 123 M A G s. For the Lichen thallus sample, there are 25 H Q M A G s, 21 M Q M A G s and a total of 46 M A G s. For the Seawater 1 sample, there are 53 H Q M A G s, 68 M Q M A G s and a total of 121 M A G s. For the Seawater 2 sample, there are 41 H Q M A G s, 82 M Q M A G s and a total of 123 M A G s. For the Hot spring sediment sample, there are 86 H Q M A G s, 35 M Q M A G s and a total of 121 M A G s. For the Wetland soil sample, there are 563 H Q M A G s, 669 M Q M A G s

and a total of 1232 M A G s. For the Grassland soil 1 sample, there are 133 H Q M A G s, 264 M Q M A G s and a total of 397 M A G s. For the Grassland soil 2 sample, there are 206 H Q M A G s, 206 M Q M A G s and a total of 412 M A G s. For the Sheep gut sample, there are 579 H Q M A G s, 457 M Q M A G s and a total of 1036 M A G s. For the Chicken gut sample, there are 269 H Q M A G s, 283 M Q M A G s and a total of 552 M A G s. For the Human gut 1 sample, there are 90 H Q M A G s, 121 M Q M A G s and a total of 211 M A G s. For the Human gut 2 sample, there are 85 H Q M A G s, 94 M Q M A G s and a total of 179 M A G s. For the Human gut (Zymo D6322) sample, there are 174 H Q M A G s, 188 M Q M A G s and a total of 362 M A G s. For the Human gut (Zymo D6331) sample, there are 177 H Q M A G s, 201 M Q M A G s and a total of 378 M A G s.

Navigate back to Figure 8.7

Description of Figure 8.8



The x-axis represents the total data in G b, ranging from 0 to 300 in increments of 50. The y-axis shows the number of M A G s, ranging from 0 to 1200 in increments of 200. The graph includes data for two sample types: soil represented in green and human gut represented in orange. The soil samples are further categorized into wetland soil, grassland soil 1 and grassland soil 2. The human gut sample is labeled as Zymo D6323. Data points and trend lines for each sample type: 1. Wetland soil: Shows a steep upward trend which is straight line, with data points starting at approximately (5, 25) and ends at (195, 1225). 2. Grassland soil 2: Displays a moderate upward trend which is a straight line with data points starting at (20, 25) and ends at (150, 425). 3. Grassland soil 1: Exhibits the lowest trend among soil samples which also forms a straight line, with data points starting at (55, 20) and ending at (275, 400). 4. Zymo D6323 (human gut): starts at nearly (0, 50), increases rapidly to (25, 300) and increases slight after that till (255, 550). The graph also

includes vertical dashed lines labeled as '1 Cell', '2 Cells', '3 Cells' and '4 Cells', indicating cell count thresholds. A legend in the top left corner identifies the sample types.

Navigate back to Figure 8.8

CHAPTER 9

Genome centric metagenomics: concepts and resources

Mick Arnold Cornelis Adriaansens and Craig William Herbold, Te Kura Pūtaiao Koiora—School of Biological Sciences, Te Whare Wnanga O Waitaha —University of Canterbury, Ōtautahi, Christchurch, New Zealand

Abstract

Affordable sequencing and advanced computational tools have made it relatively easy to generate thousands of metagenome-assembled genomes (MAGs) from metagenomic datasets, revealing a staggering level of microbial diversity. However each computational step requires several assumptions and choices, which can impact the MAG quality. Despite their usefulness in uncovering microbial diversity and community-led efforts to ensure quality MAGs often contain sequencing errors and contamination. This limitation affects how they can be interpreted. Rather than representing a single organism MAGs likely reflect a minimal, core version of a complex environmental population. The continued refinement of sequencing and computational methods holds promise for improving the accuracy and completeness of MAGs, ultimately enhancing our understanding of microbial ecosystems.

Keywords

Metagenomics; MAG; taxonomy; classification; assembly; binning; genome-centric

Advanced sequencing techniques and the rapid development of computational tools have democratized the ability to construct individual genomes from shotgun metagenomic data (Albertsen, 2023; Delmont et al., 2022). These genomes, known as metagenome-assembled genomes, or MAGs (Hugerth et al., 2015), can be processed using standard genomic tools to infer functional characteristics, which has significantly expanded the field of genome-centric metagenomics (Wilmes et al., 2009). The ease of reconstructing genomes from environmental DNA sequences has led to a rapid increase in the number of publicly available prokaryotic genomes (**Fig. 9.1**), necessitating parallel advances in methods, standards of use, and an adaptable structure for naming and organizing the vast diversity recovered (Chaumeil et al., 2019; Hedlund et al., 2022; Meyer et al., 2022; Sczyrba et al., 2017). Modern scientific discourse is rife with competing opinions regarding the utility of MAGs. This chapter is meant to introduce concepts, tools, and resources to build a fundamental understanding of prokaryotic MAGs. Although we reference a subset of available tools, this is a rapidly evolving field with continuous advancements. Readers are encouraged to consult contemporary reviews (Eren & Banfield, 2024; Kim et al., 2024) and follow the latest recommendations from community-led consortia dedicated to maintaining high scientific quality standards (Bowers et al., 2017; Meyer et al., 2022; Sczyrba et al., 2017).

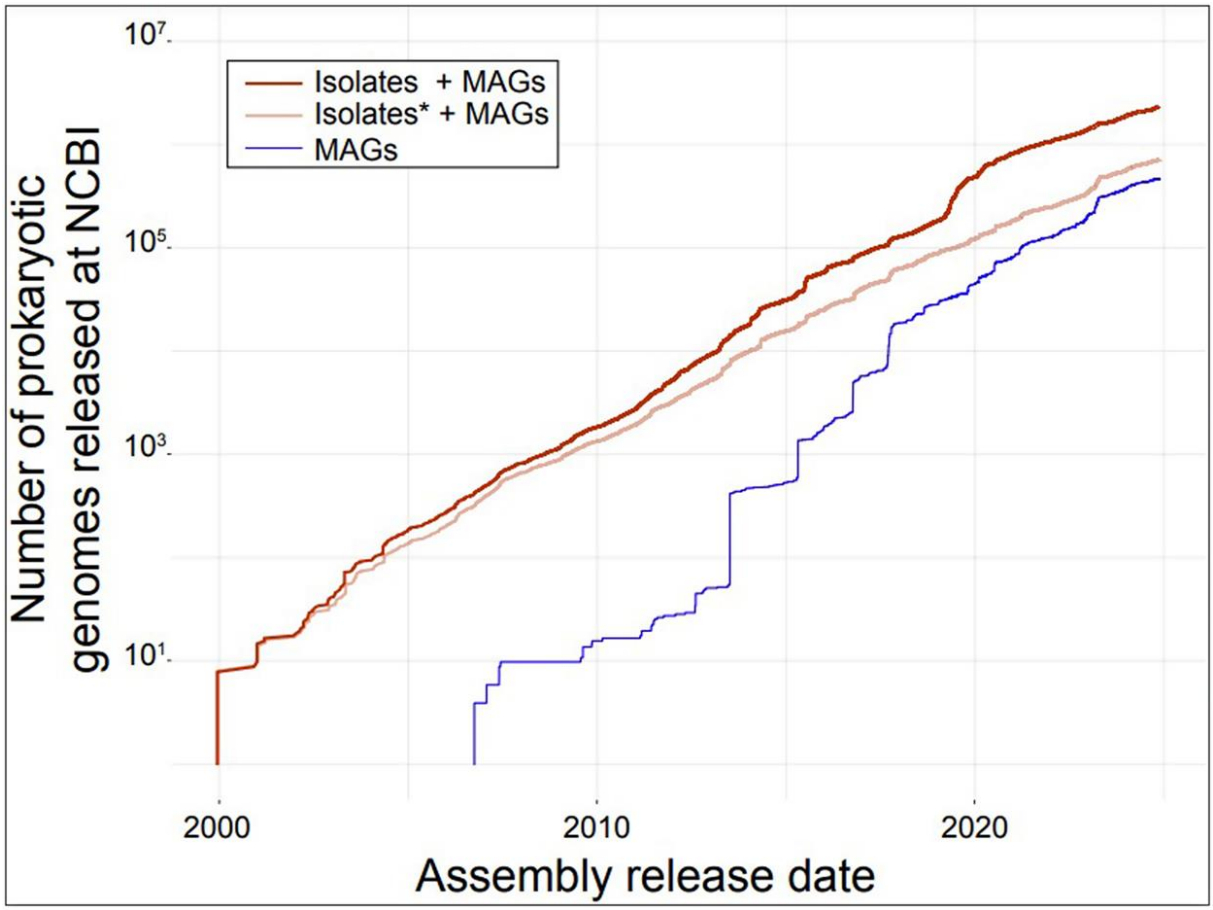


FIGURE 9.1 The increase in MAGs over time. The cumulative number of genome assemblies released by NCBI as a function of time. Total isolates + MAGs in red and MAGs alone in blue. Total isolates* + MAGs refers to the total when the 20 genera most overrepresented in genomic data are excluded (*Salmonella*, *Escherichia*, *Staphylococcus*, *Streptococcus*, *Campylobacter*, *Klebsiella*, *Listeria*, *Pseudomonas*, *Enterococcus*, *Neisseria*, *Acinetobacter*, *Shigella*, *Vibrio*, *Clostridioides*, *Bacillus*).

Follow for extended description

What is a metagenome-assembled genome?

A metagenome-assembled genome (MAG) represents a genome from a single species variant that is derived by binning metagenomic Page 192 sequence data. There is no definitive method for binning, but internally consistent

characteristics such as tetranucleotide signatures and estimated abundance form the basis for many tools in use today (Alneberg et al., 2014; Kang et al., 2019; Wu et al., 2016). The methods and data are far from perfect; therefore quality metrics typically assess how well a given MAG represents a single, complete genomic sequence from the environment (Bowers et al., 2017; Pride et al., 2003; Sczyrba et al., 2017).

How do you construct a metagenome-assembled genome?

The assembly of MAGs is largely automated and easily organized into a pipeline. Raw metagenomic data undergo sequential steps of raw read processing, assembly, binning, refining, and quality checks to reconstruct MAGs. Many automated binning tools calculate tetranucleotide composition and abundance for each contig to derive a pairwise distance. Shorter distances indicate a higher probability of contigs Page 193 that represent parts of the same genome and that can be clustered together to form a MAG (Alneberg et al., 2014; Kang et al., 2015; Kang et al., 2019; Wu et al., 2016). More complex strategies incorporate assembly graphs (Quince et al., 2021), rely on artificial intelligence (Nissen et al., 2021), or use an ensemble approach that combines binning from multiple samples (Sieber et al., 2018) or multiple binning software (Uritskiy et al., 2018). More recent tools even combine all these approaches (Qiu et al., 2024).

Ultimately binning methods rely on the quality of assembled metagenomic data, which affects MAG quantity and quality (Gerner et al., 2018; Sczyrba et al., 2017). Contigs in a MAG are small subsamples of a larger genome. Just as small sample sizes show larger variation than large sample sizes, short contigs show more variation than long contigs (**Fig. 9.2**). Given this inherent variance, fragmented metagenomic assemblies do not typically lead to high-quality MAGs. Page 194 Fragmentation can result from population-level variation in single nucleotides or genome structure; thus complex, diverse populations are less likely to be represented by MAGs (Sczyrba et al., 2017).

Compounding this are other inherent features known to cause problems for single-isolate assemblies, such as the presence of ribosomal RNA operons, CRISPR, or transposable elements among others (Wick et al., 2023). Restricting input to contigs greater than 2 kb in length improves binning performance (Dick et al., 2009), but also inherently excludes much of a fragmented metagenomic assembly.

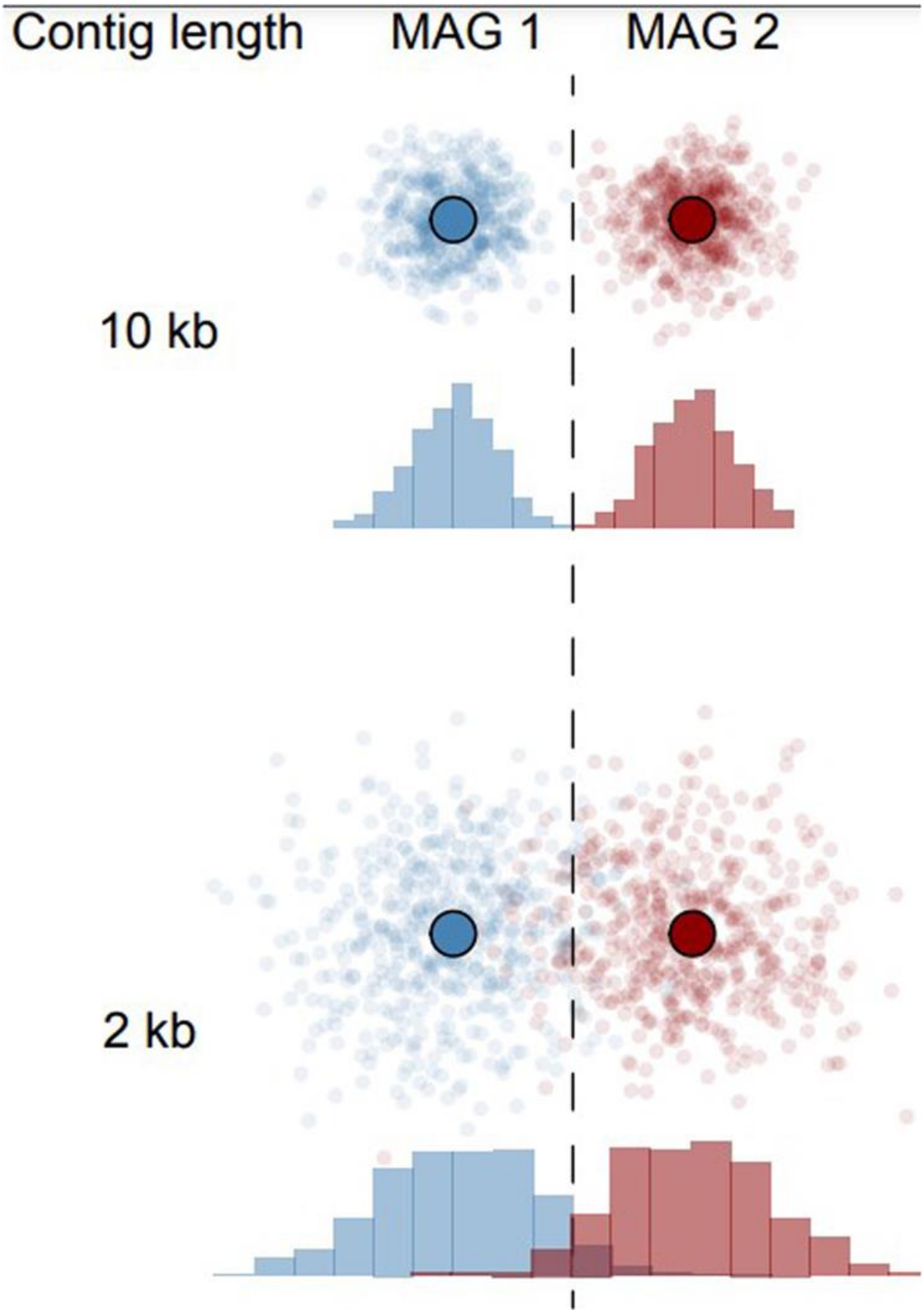


FIGURE 9.2 Effect of contig length on binning. Contigs (semitransparent dots) are plotted around the true

genome-wide average (opaque circles) of two MAGs in a theoretical 2D tetranucleotide space. Longer contigs (top) show low variance and more closely resemble the genome-wide average. The tighter distribution around the genome-wide average allows the two MAGs to be separated into their appropriate bin (dotted line). The greater variance of the short contigs (bottom) results in higher variance and incorrect binning.

Follow for extended description

Because contig length is such an important factor for binning, it is essential to consider the downstream effects of sequencing and assembly choices. Long-read sequencing technologies such as Oxford Nanopore Promethion R10.4 and Pacific Biosciences HiFi generate long contig-length reads (>10 kb) that can span repeat regions to aid in assembly. These reads have error rates of ~1%–2%, the majority of which are indels (Hon et al., 2020; Sereika et al., 2022). One way of reducing this error is through polishing them prior to assembly with high-quality short-read data (Wick et al., 2023). In the absence of short-read data, errors can be reduced through assembly. Several reports have shown that binning assembled long-read data can produce complete, circularized MAGs (cMAGs) of abundant organisms (Albertsen, 2023; Kim et al., 2022; Moss et al., 2020; Sereika et al., 2022; Singleton et al., 2021). When using MAGs constructed from long reads, it is important to consider that error correction results from read depth. Low-coverage regions of a MAG binned with long reads are prone to indel errors, which manifest as frameshifts in predicted proteins. Global average stats do not identify those regions, so these MAGs may not be as perfect as they seem.

Despite the advantages of long reads there are many cases for which their generation is unfeasible. Large quantities of high-purity, high molecular weight DNA are required to produce libraries for long-read sequencing; however many environmental samples require harsh DNA extraction

protocols which produce low quantities of highly sheared DNA (Leff et al., 1995). Fragments of 400–500 bases are optimal for short-read sequencing (Tan et al., 2019) making it more suitable in these cases. The most common short-read sequencing platforms, Illumina and MGI, produce large volumes of reads exceeding 99% accuracy, with miscalled bases as the most common type of error (Hu et al., 2024; Stoler & Nekrutenko, 2021). The ease of generating sequencing libraries makes short-read sequencing useful for generating thousands of MAGs *en masse* from diverse environments and Page 195 estimating their environmental abundances (Nishimura & Yoshizawa, 2022; Parks et al., 2017; Xie et al., 2021). However these short fragments do not span repeated elements and therefore MAGs from short-read data tend to be made of hundreds, if not thousands of small contigs, some proportion of which are likely improperly binned (Orakov et al., 2021).

Metagenome-assembled genome quality standards

The probabilistic nature of binning requires a balance between type I and type II error, resulting in MAGs that are both contaminated and not complete. Quality assessment criteria have been developed to quantify these two characteristics using single-copy markers, which are proteins that are consistently found in single copies across a set of examined genomes. While universal single-copy markers exist that are useful for calculating deep phylogenomic relationships between all major taxonomic groups (Wu & Eisen, 2008), they tend to be ribosomal proteins and other informational genes (Chklovski et al., 2023; Rivera et al., 1998) that make up a small proportion of any given genome (Dagan & Martin, 2006). Furthermore some subsets of ribosomal proteins are organized into operons in prokaryotic genomes (Siefert et al., 1997; Tirumalai et al., 2023). The presence or absence of such a small number of co-localized markers is not as useful as a larger set of markers that is distributed more evenly across the genome. CheckM

(hereafter CheckM1) overcame this challenge by defining lineage-specific sets of single-copy proteins (Parks et al., 2015).

In CheckM1 a small universal set of 43 hidden Markov models (HMMs) representing 34 single-copy proteins from available genomes and MAGs was used to generate a reference tree of thousands of isolate genomes and MAGs. This tree was used to define well-supported clades as "lineages." Each lineage was then examined for a set of proteins that could be found exactly one time in nearly all members of the clade. CheckM1 calculates the quality of a novel MAG by first identifying and aligning the 43 phylogenetically informative markers using HMMs. This alignment is used to phylogenetically place the MAG into the reference tree to determine which lineage-specific marker set to use. MAG completeness is the proportion of unique lineage-specific markers observed at least one time. Contamination is the proportion of observed unique lineage-specific markers that were observed multiple times. CheckM1 included an interesting secondary calculation "strain

heterogeneity" which is a measure of the proportion of contamination that might be due to overbinning very closely related taxa. When interpreting results keep in mind that it is an inexact estimate. CheckM1 can return values as low as 95% complete and up to 5% contaminated for complete, circularized isolate genomes (Parks et al., 2015).

Despite its widespread use the reference data in checkM V1 have not been updated since 2015. Many "microbial dark matter" lineages have been discovered and/or reorganized since that time (Brown et al., 2015; Parks et al., 2018; Rinke et al., 2021). These new lineages are not part of the reference tree, undermining the goal of using lineage-specific markers. CheckM2 overcomes these shortcomings by utilizing machine learning (Chklovski et al., 2023). Neural networks and gradient-boosted decision trees were trained using a diverse assortment of genomes and MAGs to predict completeness and contamination of prokaryotic MAGs. On average it was shown to predict quality more accurately than checkM1 and its predictive models can be updated more easily, suggesting that it will continue to be fit for purpose for

the foreseeable future. Like CheckM1, CheckM2 is an inexact estimator of completeness and contamination. However they use different methodology and therefore can be seen as independent estimators from one another. Because of this the best course of action is to use both tools to get a better estimate of quality (Chklovski et al., 2023).

The assessment of MAGs by CheckM1 and CheckM2 is based on protein-coding genes, which ensure that MAGs are suitable for metabolic inference and phylogenomics. However these tools overlook rRNA and tRNA genes and could easily report a MAG that is missing rRNA as 100% complete. Because rRNA genes are found in multiple copies and are longer than short-read fragments, they tend not to assemble well. Even when assembled they carry a tetranucleotide signature that differs from the rest of the genome and are therefore rarely binned into short-read based MAGs (Song et al., 2022).

The Genomic Standards Consortium established an extensive list of characteristics for MAGs (as well as single-cell amplified genomes—SAGs) to define quality ranging from "low-quality draft" to "finished" (Bowers et al., 2017). Low- and medium-quality MAGs are defined by completeness and contamination estimates exclusively based on single-copy marker genes. For a MAG to be considered high quality it must exceed 90% completeness, have less than 10% contamination, and contain 23, 16 and 5 S ribosomal RNA genes and tRNAs for at least 18 amino acids. For a MAG to be considered "finished" it must be a Page 197 complete, circularized contig with no gaps or ambiguities and a consensus quality score of Q50 or better. Because long-read sequences are long enough to span the entire rRNA operon, MAGs generated from long-read datasets are more likely to meet these more stringent thresholds (Albertsen, 2023; Hon et al., 2020; Sereika et al., 2022).

Naming and organizing the uncultivated

The microbial diversity uncovered through MAGs has reshaped our understanding of microbial diversity and relationships between major eukaryotic groups (Hug et al., 2016; Spang et al., 2015). A taxonomic structure

that can accommodate this diversity has been called for by the scientific community (Hugenholtz et al., 2021; Murray et al., 2020). However official taxonomic designations are governed by various nomenclatural codes, which consistently require physical specimens to act as "type" material for establishing a genus and species. The International Code of Nomenclature of Prokaryotes (ICNP) has as recently as 2022 rejected the notion that sequence (i.e., MAGs) can act as type, rendering entire branches of the tree-of-life devoid of official taxonomic classification (Murray et al., 2020; Oren et al., 2023). In response a parallel system of nomenclature has been developed, the SeqCode, which provides guidelines for naming uncultivated organisms while also retaining the majority of the rules set forth by the ICNP in the hope that eventually the two codes will merge (Hedlund et al., 2022).

SeqCode provides a mechanism to propose genus and species names for MAGs, the genus name of which can lend its root to the names of higher ranks, up to phylum (Parker et al., 2019). For the SeqCode the type material is the MAG itself, which should be>90% complete (Parks et al., 2022) and<5% contaminated (Whitman et al., 2024) but given the limitations discussed above in assessing quality with checkM1/2 special situations may be dealt with on a case-by-case basis. In addition to meeting these basic requirements the MAG, as well as the raw metagenomic data from which it was derived, must be publicly accessible in an International Nucleotide Sequence Database Collaboration database. The MAG must be compared to types that have already been recognized by SeqCode to determine whether it represents a new species and/or genus. Finally the MAG should be presented in an effective publication, which describes the MAG and the methods used to derive it, which ensures that peer-review has occurred (Whitman et al., 2024). Page 198

At the same time that the SeqCode provides a mechanism for naming MAGs and is committed to following the ICNP rules for propagating names to higher ranks, the genome taxonomy database (GTDB) has emerged as a de facto authority on taxonomy for the scientific community (Parks et al., 2022).

Before GTDB it was left to individual researchers to determine at what level a MAG represented a novel lineage (e.g., family and order). GTDB is regularly updated with newly discovered isolate genomes and MAGs and re-evaluates membership to higher-level taxonomy with each release.

The core of the GTDB structure is type genomes. To determine type genomes the GTDB clusters all available Genbank and Refseq prokaryotic genomes by average nucleotide identity (ANI) to build a set of species-level clusters and picks the one with the highest score. Scores are based on several criteria, the highest weighted metric of which is whether the genome is from an isolated type strain (Parks et al., 2022). After that quality metrics are considered, along with consistency in ANI values with other members of the cluster, and a continuity criterion, which requires that a new genome be significantly better in quality to replace an existing type genome (Parks et al., 2022).

For each update the GTDB uses type genomes to build phylogenomic trees that define taxonomic clades. A concatenated alignment of curated single-copy markers that are common to most bacteria is used to build a reference bacterial phylogenomic tree (Parks et al., 2018) and a concatenated alignment curated single-copy markers that are common to most archaea is used to build a reference archaeal phylogenomic tree (Rinke et al., 2021). Each tree is rooted near the midpoint and branch lengths are scaled across the tree so that the distance from root to each tip is exactly 1 (Parks et al., 2018). These scaled lengths are then used to assign a relative evolutionary divergence (RED) value to each node in the rooted tree and define taxonomically informative nodes. The deepest taxonomically informative nodes in the tree correspond to the base of phylum-level groups, with low RED values. The most shallow taxonomically informative nodes in the tree correspond to genus-level groups with high RED values, thus ensuring that taxonomy is phylogenetically informed.

An underlying premise of the GTDB is that equal ranks should be defined so that they contain a similar amount of diversity. Before its development

this was not true (Parks et al., 2018). However, when applied to the GTDB reference tree, this scaling and normalized ranking procedure has resulted in an impressively stable taxonomy over time. Page 199 Major rearrangements are rare at this point, and the result is an organizational structure that has and will continue to profoundly inform our understanding of microbial diversity (Chuvochina et al., 2023; Dutkiewicz et al., 2024; Waite et al., 2017; Waite et al., 2020).

GTDB-Tk was developed as a tool to classify novel MAGs into the existing GTDB taxonomic framework (Chaumeil et al., 2019). For each MAG a complete lineage is returned along with a RED value estimate for how deeply the MAG attaches to the reference tree (Chaumeil et al., 2019). The mechanism behind this placement is very similar to the way that checkM works. Bacterial or archaeal markers are identified from a novel MAG and aligned into an HMM-based reference alignment. MAGs are then placed into the tree at the most likely position, which provides the estimated RED value and taxonomy. With this tool MAGs can routinely be assessed for their taxonomic affiliation and whether they likely represent a new lineage, combined with the existing classification database.

What are key differences between a metagenome-assembled genome and an isolate genome?

Despite the plethora of tools to make the best MAGs and their integration with isolate genomes into a unified phylogenomic structure MAGs fundamentally represent a different type of genomic data than an isolate genome. A genome from an isolate is a singular, clonal representative. A MAG is instead a hypothetical reconstruction superficially resembling an isolate genome. In reality it is a smoothed out version of the variation found in a natural population (Charuvaka & Rangwala, 2011).

Assemblies are constructed through consensus by treating low-frequency observations as noise and may not reflect natural sequence precisely

(Rahman & Medvedev, 2022). It has also been shown that for a single species complex, MAGs can omit much of the population core and variable genome (**Fig. 9.3**), despite appearing high quality due to the ease at which single-copy markers are binned (Meziti et al., 2021). At the same time MAGs are notorious for containing undetected contamination (Astashyn et al., 2024; Mattock & Watson, 2023; Orakov et al., 2021; Shaiber & Eren, 2019).

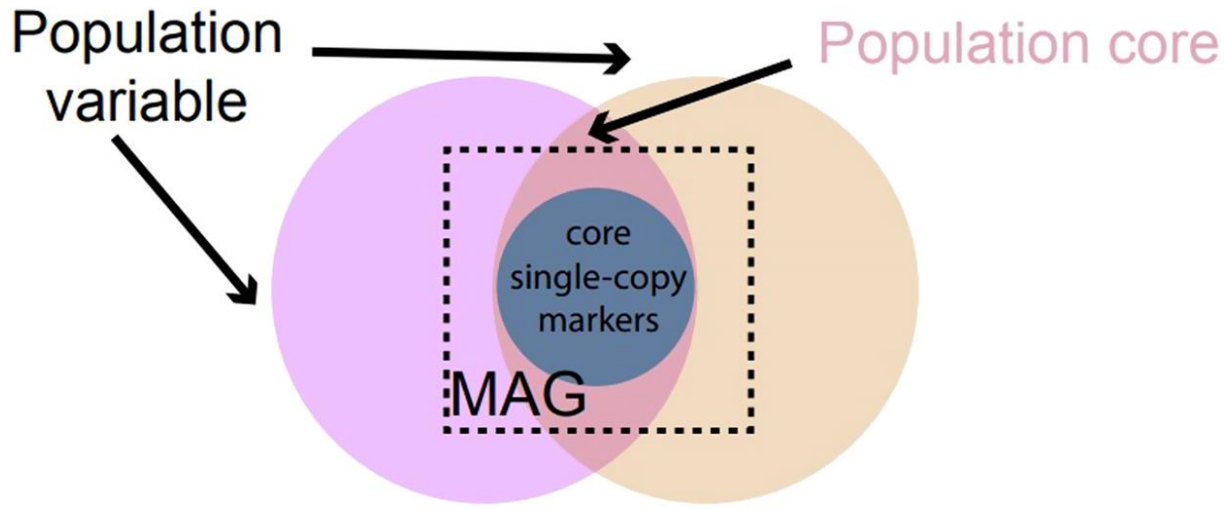


FIGURE 9.3 MAG representation of a population. Two variants of a single species are represented as circles of a Venn diagram. The overlap between them represents the population core. A MAG constructed from metagenomic data with both variants will likely contain the majority of the population core but less of the variable genome as per Meziti et al. (2021). Co-assembled contigs (blue inner circle) are a stable subset of the population core that includes the single-copy markers used for quality assessment.

Follow for extended description

Given these complications it seems unlikely that any single MAG is an exact replica of an environmental genome. But should this stop us from using them? One typical use of MAGs is for classification and discussions on

diversity. As long as a MAG contains all or most Page 200 single-copy markers used for phylogenomic analysis, it can be classified. Even minor levels of contamination do not seem to affect this (Mussig et al., 2024). Another typical use of a MAG is in elucidating core metabolism using the presence and absence of genes that are indicative of a metabolic pathway of interest. If a metabolic process is relevant to the species as a whole, it is likely contained within the recovered MAG core.

Conclusions and outlook

Advanced computational techniques and affordable sequencing have combined to provide a relatively easy and automated way to generate MAGs from metagenomic datasets. MAGs in turn have revolutionized our view of overall microbial diversity and greatly expanded the tree of life. This massive change in the way we view life has necessitated the establishment of new tools, standards, and a community-led framework for organizing it into a meaningful taxonomy. Although it is tempting to interpret MAGs and isolate genomes in the same way, there is reason for caution. MAGs are constructed using techniques that are imprecise and introduce error. As a rule they should be regarded as incomplete, contaminated representations of a population, rather than as an individual cell. Despite this they are well-suited for phylogenomics and core metabolic construction. Many of the assembly and binning issues that affect MAG quality may be addressed through the use of long-read sequencing. If long-read sequencing accuracy continues to increase and sample requirements continue to be loosened, Page 201 we may soon be producing thousands of complete, circularized MAGs from environmental samples.

MAGs are regularly constructed for viruses and eukaryotes (Delmont et al., 2022; Johansen et al., 2022), and BUSCO, a tool for their assessment using the same marker-based strategy as CheckM1, has been developed (Simão et al., 2015; Tegenfeldt et al., 2024). While we did not discuss these MAGs, the concepts introduced here are general and should provide a suitable

knowledge base from which to expand into these specialized topics. The study of eukaryotic MAGs in particular is still in its infancy, and it will be interesting to see how the field develops over time. If it is as impactful for microbial eukaryotes as it has been for prokaryotes, we may see yet another revolution in our understanding of diversity of life on Earth.

Acknowledgments

MACA is supported by the University of Canterbury Faculty of Science Doctoral Scholarship.

References

Albertsen, 2023 Albertsen M. Long-read metagenomics paves the way toward a complete microbial tree of life. *Nature methods*. 2023;20:30–31 **https://doi.org/10.1038/s41592-022-01726-6**.

Alneberg et al., 2014 Alneberg J, Bjarnason BS, De Bruijn I, et al. Binning metagenomic contigs by coverage and composition. *Nature methods*. 2014;11(11):1144–1146 **https://doi.org/10.1038/nmeth.3103**.

Astashyn et al., 2024 Astashyn A, Tvedte ES, Sweeney D, et al. Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome biology*. 2024;25:60 **https://doi.org/10.1186/s13059-024-03198-7**.

Bowers et al., 2017 Bowers RM, Kyrpides NC, Stepanauskas R, et al.... Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology*. 2017;35:725–731 **https://doi.org/10.1038/nbt.3893**.

Brown et al., 2015 Brown CT, Hug LA, Thomas BC, et al. Unusual biology across a group comprising more than 15% of domain

Bacteria. *Nature*. 2015;523(7559):208–211 **https://doi.org/10.1038/nature14486**.

Charuvaka and Rangwala, 2011 Charuvaka A, Rangwala H. Evaluation of short read metagenomic assembly. *BMC Genomics*. 2011;12(Suppl 2):S8 **https://doi.org/10.1186/1471-2164-12-S2-S8**.

Chaumeil et al., 2019 Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*. 2019;36:1925–1927 **https://doi.org/10.1093/bioinformatics/btz848**.

Chklovski et al., 2023 Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. CheckM2: A rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nature methods*. 2023;20:1203–1212 **https://doi.org/10.1038/s41592-023-01940-w**.

Chuvochina et al., 2023 Chuvochina M, Mussig AJ, Chaumeil PA, et al. Proposal of names for 329 higher rank taxa defined in the Genome Taxonomy Database under two prokaryotic codes. *FEMS Microbiology Letters*. 2023;370 **https://doi.org/10.1093/femsle/fnad071**.

Dagan and Martin, 2006 Dagan T, Martin W. The tree of one percent. *Genome biology*. 2006;7 **https://doi.org/10.1186/gb-2006-7-10-118**.

Delmont et al., 2022 Delmont TO, Gaia M, Hinsinger DD, et al.... Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics*. 2022;2 **https://doi.org/10.1016/j.xgen.2022.100123**.

Dick et al., 2009 Dick GJ, Andersson AF, Baker BJ, et al. Community-wide analysis of microbial genome sequence signatures. *Genome biology*. 2009;10(8):R85 **https://doi.org/10.1186/gb-2009-10-8-r85**.

Dutkiewicz et al., 2024 Dutkiewicz Z, Singleton CM, Sereika M, et al. Proposal of Patescibacterium danicum gen nov., sp nov in the ubiquitous ultrasmall bacterial phylum Patescibacteriota phyl nov. *bioRxiv*. 2024;5(1):ycae147 **https://doi.org/10.1101/2024.11.01.618662**.

Eren and Banfield, 2024 Eren AM, Banfield JF. Modern microbiology: Embracing complexity through integration across scales. *Cell*. 2024;187(19):5151–5170 **https://doi.org/10.1016/j.cell.2024.08.028**.

Gerner et al., 2018 Gerner SM, Rattei T, Graf AB. Assessment of urban microbiome assemblies with the help of targeted in silico gold standards. *Biology direct*. 2018;13:22 **https://doi.org/10.1186/s13062-018-0225-6**.

Hedlund et al., 2022 Hedlund BP, Chuvochina M, Hugenholtz P, et al. SeqCode: A nomenclatural code for prokaryotes described from sequence data. *Nature Microbiology*. 2022;7(10):1702–1708 **https://doi.org/10.1038/s41564-022-01214-9**.

Hon et al., 2020 Hon T, Mars K, Young G, et al. Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data*. 2020;7 **https://doi.org/10.1038/s41597-020-00743-4**.

Hu et al., 2024 Hu T, Chen J, Lin X, et al. Comparison of the DNBSEQ platform and Illumina HiSeq 2000 for bacterial genome assembly. *Scientific Reports*. 2024;14 **https://doi.org/10.1038/s41598-024-51725-0**.

Hug et al., 2016 Hug LA, Baker BJ, Anantharaman K, et al. A new view of the tree of life. *Nature Microbiology*. 2016;1 **https://doi.org/10.1038/nmicrobiol.2016.48**.

Hugenholtz et al., 2021 Hugenholtz P, Chuvochina M, Oren A, Parks DH, Soo RM. Prokaryotic taxonomy and nomenclature in the age

of big sequence data. *ISME Journal*. 2021;15(7):1879–1892 **https://doi.org/10.1038/s41396-021-00941-x**.

Hugerth et al., 2015 Hugerth LW, Larsson J, Alneberg J, et al. Metagenome-assembled genomes uncover a global brackish microbiome. *Genome biology*. 2015;16 **https://doi.org/10.1186/s13059-015-0834-7**.

Johansen et al., 2022 Johansen J, Plichta DR, Nissen JN, et al. Genome binning of viral entities from bulk metagenomics data. *Nature Communications*. 2022;13 **https://doi.org/10.1038/s41467-022-28581-5**.

Kang et al., 2015 Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*. 2015;3:e1165 **https://doi.org/10.7717/peerj.1165**.

Kang et al., 2019 Kang DD, Li F, Kirton E, et al. MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*. 2019;7(7):e7359 **https://doi.org/10.7717/peerj.7359**.

Kim et al., 2022 Kim CY, Ma J, Lee I. HiFi metagenomic sequencing enables assembly of accurate and complete genomes from human gut microbiota. *Nature Communications*. 2022;13:6367 **https://doi.org/10.1038/s41467-022-34149-0**.

Kim et al., 2024 Kim N, Ma J, Kim W, Kim J, Belenky P, Lee I. Genome-resolved metagenomics: A game changer for microbiome medicine. *Experimental & Molecular Medicine*. 2024;56(7):1501–1512 **https://doi.org/10.1038/s12276-024-01262-7**.

Leff et al., 1995 Leff LG, Dana JR, McArthur JV, Shimkets LJ. Comparison of methods of DNA extraction from stream sediments. *Applied and Environmental Microbiology*.

1995;61(3):1141–1143 **https://doi.org/10.1128/aem.61.3.1141-1143.1995**.

Mattock and Watson, 2023 Mattock J, Watson M. A comparison of single-coverage and multi-coverage metagenomic binning reveals extensive hidden contamination. *Nature methods*. 2023;20(8):1170–1173 **https://doi.org/10.1038/s41592-023-01934-8**.

Meyer et al., 2022 Meyer F, Fritz A, Deng ZL, et al.... Critical assessment of metagenome interpretation: The second round of challenges. *Nature methods*. 2022;19(4):429–440 **https://doi.org/10.1038/s41592-022-01431-4**.

Meziti et al., 2021 Meziti A, Rodriguez-R LM, Hatt JK, Peña-Gonzalez A, Levy K, Konstantinidis KT. The reliability of metagenome-assembled genomes (MAGs) in representing natural populations: Insights from comparing MAGs against isolate genomes derived from the same fecal sample. *Applied and Environmental Microbiology*. 2021;87(6):1–15 **https://doi.org/10.1128/AEM.02593-20**.

Moss et al., 2020 Moss EL, Maghini DG, Bhatt AS. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nature Biotechnology*. 2020;38:701–707 **https://doi.org/10.1038/s41587-020-0422-6**.

Murray et al., 2020 Murray AE, Freudenstein J, Gribaldo S, et al.... Roadmap for naming uncultivated archaea and bacteria. *Nature Microbiology*. 2020;5(8):987–994 **https://doi.org/10.1038/s41564-020-0733-x**.

Mussig et al., 2024 Mussig AJ, Chaumeil PA, Chuvochina M, Rinke C, Parks DH, Hugenholtz P. Putative genome contamination has minimal impact on the GTDB taxonomy. *Microbiology Society, Australia Microbial Genomics*. 2024;10 **https://doi.org/10.1099/mgen.0.001256**.

Nishimura and Yoshizawa, 2022 Nishimura Y, Yoshizawa S. The OceanDNA MAG catalog contains over 50,000 prokaryotic genomes originated from various marine environments. *Scientific Data*. 2022;9 **https://doi.org/10.1038/s41597-022-01392-5**.

Nissen et al., 2021 Nissen JN, Johansen J, Allesøe RL, et al. Improved metagenome binning and assembly using deep variational autoencoders. *Nature Biotechnology*. 2021;39(5):555–560 **https://doi.org/10.1038/s41587-020-00777-4**.

Orakov et al., 2021 Orakov A, Fullam A, Coelho LP, et al. GUNC: Detection of chimerism and contamination in prokaryotic genomes. *Genome biology*. 2021;22 **https://doi.org/10.1186/s13059-021-02393-0**.

Oren et al., 2023 Oren A, Arahal DR, Göker M, Moore ERB, Rossello-Mora R, Sutcliffe IC. International code of nomenclature of prokaryotes Prokaryotic code (2022 revision). *International Journal of Systematic and Evolutionary Microbiology*. 2023;73 **https://doi.org/10.1099/ijsem.0.005585**.

Parker et al., 2019 Parker CT, Tindall BJ, Garrity GM. International code of nomenclature of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology* 2019;.

Parks et al., 2015 Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*. 2015;25(7):1043–1055 **https://doi.org/10.1101/gr.186072.114**.

Parks et al., 2017 Parks DH, Rinke C, Chuvochina M, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*. 2017;2(11):1533–1542 **https://doi.org/10.1038/s41564-017-0012-7**.

Parks et al., 2018 Parks DH, Chuvochina M, Waite DW, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*. 2018;36(10):996–1004 **https://doi.org/10.1038/nbt.4229**.

Parks et al., 2022 Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil PA, Hugenholtz P. GTDB: An ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*. 2022;50(1):D785 **https://doi.org/10.1093/nar/gkab776**.

Pride et al., 2003 Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Research*. 2003;13(2):145–156 **https://doi.org/10.1101/gr.335003**.

Qiu et al., 2024 Qiu Z, Yuan L, Lian CA, et al.... BASALT refines binning from metagenomic data and increases resolution of genome-resolved metagenomic analysis. *Nature Communications*. 2024;15(1):2179 **https://doi.org/10.1038/s41467-024-46539-7**.

Quince et al., 2021 Quince C, Nurk S, Raguideau S, et al. STRONG: Metagenomics strain resolution on assembly graphs. *Genome biology*. 2021;22(1):214 **https://doi.org/10.1186/s13059-021-02419-7**.

Rahman and Medvedev, 2022 Rahman A, Medvedev P. Assembler artifacts include misassembly because of unsafe unitigs and underassembly because of bidirected graphs. *Genome Research*. 2022;32(9):1746–1753 **https://doi.org/10.1101/gr.276601.122**.

Rinke et al., 2021 Rinke C, Chuvochina M, Mussig AJ, et al. A standardized archaeal taxonomy for the Genome Taxonomy

Database. *Nature Microbiology*. 2021;6(7):946–959 **https://doi.org/10.1038/s41564-021-00918-8**.

Rivera et al., 1998 Rivera MC, Jain R, Moore JE, Lake JA. Genomic evidence for two functionally distinct gene classes. *Proceedings of the National Academy of Sciences of the United States of America*. 1998;95(11):6239–6244 **https://doi.org/10.1073/pnas.95.11.6239**.

Sczyrba et al., 2017 Sczyrba A, Hofmann P, Belmann P, et al.... Critical Assessment of metagenome interpretation—A benchmark of metagenomics software. *Nature methods*. 2017;14(11):1063–1071 **https://doi.org/10.1038/nmeth.4458**.

Sereika et al., 2022 Sereika M, Kirkegaard RH, Karst SM, et al. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nature methods*. 2022;19(7):823–826 **https://doi.org/10.1038/s41592-022-01539-7**.

Shaiber and Eren, 2019 Shaiber A, Eren AM. Composite metagenome-assembled genomes reduce the quality of public genome repositories. *mBio*. 2019;10 **https://doi.org/10.1128/mBio.00725-19**.

Sieber et al., 2018 Sieber CMK, Probst AJ, Sharrar A, et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology*. 2018;3(7):836–843 **https://doi.org/10.1038/s41564-018-0171-1**.

Siefert et al., 1997 Siefert JL, Martin KA, Abdi F, Widger WR, Fox GE. Conserved gene clusters in bacterial genomes provide further support for the primacy of RNA. *Journal of Molecular Evolution*. 1997;45(5):467–472 **https://doi.org/10.1007/PL00006251**.

Simão et al., 2015 Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: Assessing genome

assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015;31(19):3210–3212 **https://doi.org/10.1093/bioinformatics/btv351**.

Singleton et al., 2021 Singleton CM, Petriglieri F, Kristensen JM, et al. Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. *Nature Communications*. 2021;12:2009 **https://doi.org/10.1038/s41467-021-22203-2**.

Song et al., 2022 Song W, Zhang S, Thomas T, Robinson P. MarkerMAG: Linking metagenome-assembled genomes (MAGs) with 16S rRNA marker genes using paired-end short reads. *Bioinformatics*. 2022;38(15):3684–3688 **https://doi.org/10.1093/bioinformatics/btac398**.

Spang et al., 2015 Spang A, Saw JH, Jørgensen SL, et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*. 2015;521(7551):173–179 **https://doi.org/10.1038/nature14447**.

Stoler and Nekrutenko, 2021 Stoler N, Nekrutenko A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics and Bioinformatics*. 2021;3(1):lqab019 **https://doi.org/10.1093/nargab/lqab019**.

Tan et al., 2019 Tan G, Opitz L, Schlapbach R, Rehrauer H. Long fragments achieve lower base quality in Illumina paired-end sequencing. *Scientific Reports*. 2019;9 **https://doi.org/10.1038/s41598-019-39076-7**.

Tegenfeldt et al., 2024 Tegenfeldt F, Kuznetsov D, Manni M, Berkeley M, Zdobnov EM, Kriventseva EV. OrthoDB and BUSCO update: Annotation of orthologs with wider sampling of genomes. *Nucleic Acids Research* 2024; **https://doi.org/10.1093/nar/gkae987**.

Tirumalai et al., 2023 Tirumalai MR, Sivaraman RV, Kutty LA, Song EL, Fox GE. Ribosomal Protein Cluster Organization in Asgard Archaea. *Archaea (Vancouver, B.C.)*. 2023;2023:5512414 **https://doi.org/10.1155/2023/5512414**.

Uritskiy et al., 2018 Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome*. 2018;6:158 **https://doi.org/10.1186/s40168-018-0541-1**.

Waite et al., 2017 Waite DW, Vanwonterghem I, Rinke C, et al. Comparative genomic analysis of the class Epsilonproteobacteria and proposed reclassification to Epsilonbacteraeota (phyl nov.). *Frontiers in Microbiology*. 2017;8:682 **https://doi.org/10.3389/fmicb.2017.00682**.

Waite et al., 2020 Waite DW, Chuvochina M, Pelikan C, et al. Proposal to reclassify the proteobacterial classes deltaproteobacteria and oligoflexia, and the phylum thermodesulfobacteria into four phyla reflecting major functional capabilities. *International Journal of Systematic and Evolutionary Microbiology*. 2020;70(11):5972–6016 **https://doi.org/10.1099/ijsem.0.004213**.

Whitman et al., 2024 Whitman WB, Chuvochina M, Hedlund BP, et al. Why and how to use the SeqCode. *mLife*. 2024;3(1):1–13 **https://doi.org/10.1002/mlf2.12092**.

Wick et al., 2023 Wick RR, Judd LM, Holt KE. Assembling the perfect bacterial genome using Oxford Nanopore and Illumina sequencing. *PLoS computational biology*. 2023;19:e1010905 **https://doi.org/10.1371/journal.pcbi.1010905**.

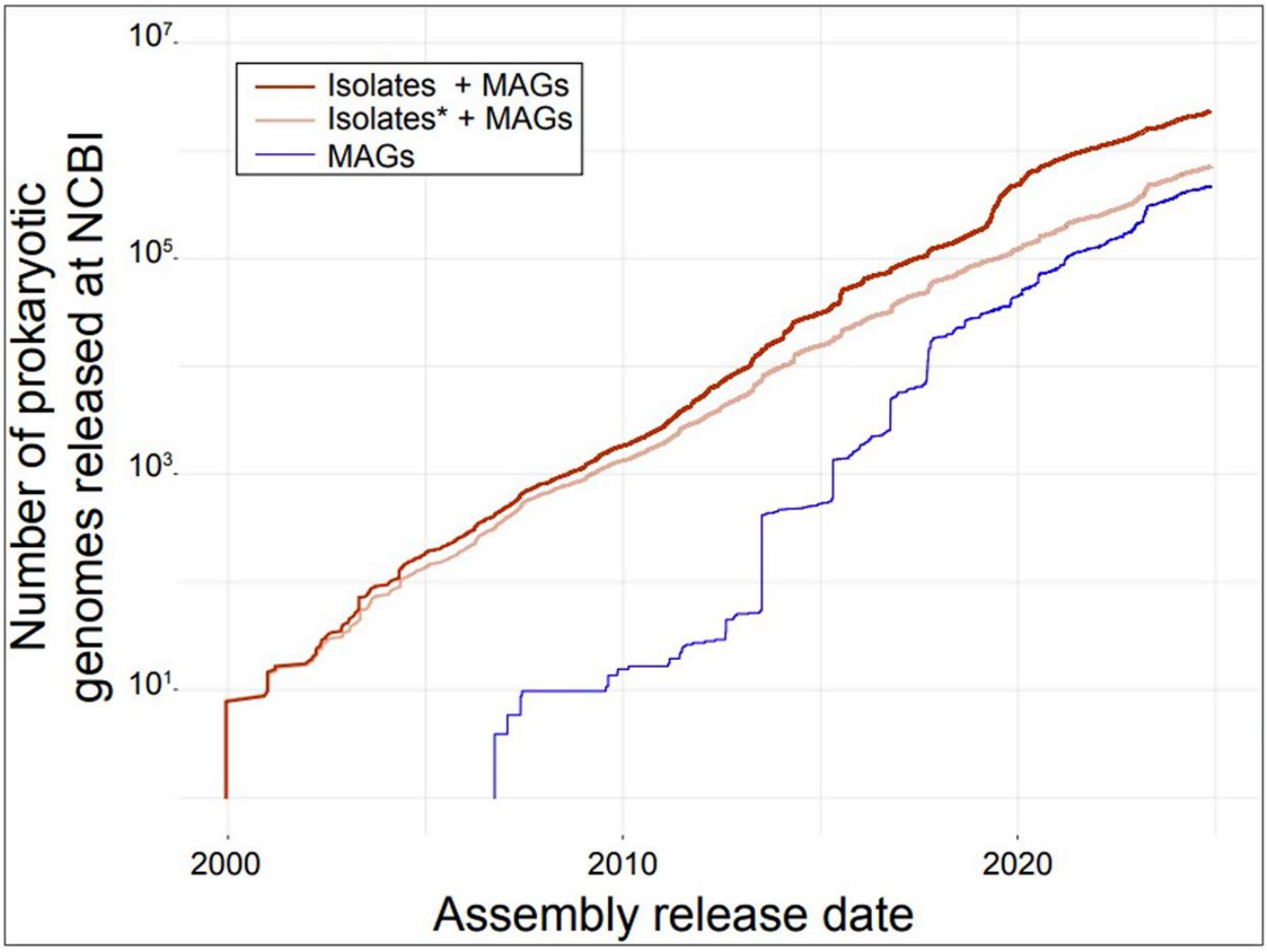
Wilmes et al., 2009 Wilmes P, Simmons SL, Denef VJ, Banfield JF. The dynamic genetic repertoire of microbial communities. *FEMS Microbiology Reviews*. 2009;33(1):109–132 **https://doi.org/10.1111/j.1574-6976.2008.00144.x**.

Wu and Eisen, 2008 Wu M, Eisen JA. A simple, fast, and accurate method of phylogenomic inference. *Genome biology*. 2008;9(10):R151 **https://doi.org/10.1186/gb-2008-9-10-r151**.

Wu et al., 2016 Wu Y-W, Simmons BA, Singer SW. MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*. 2016;32:605–607 **https://doi.org/10.1093/bioinformatics/btv638**.

Xie et al., 2021 Xie F, Jin W, Si H, et al. An integrated gene catalog and over 10,000 metagenome-assembled genomes from the gastrointestinal microbiome of ruminants. *Microbiome*. 2021;10(1):228.

Description of Figure 9.1

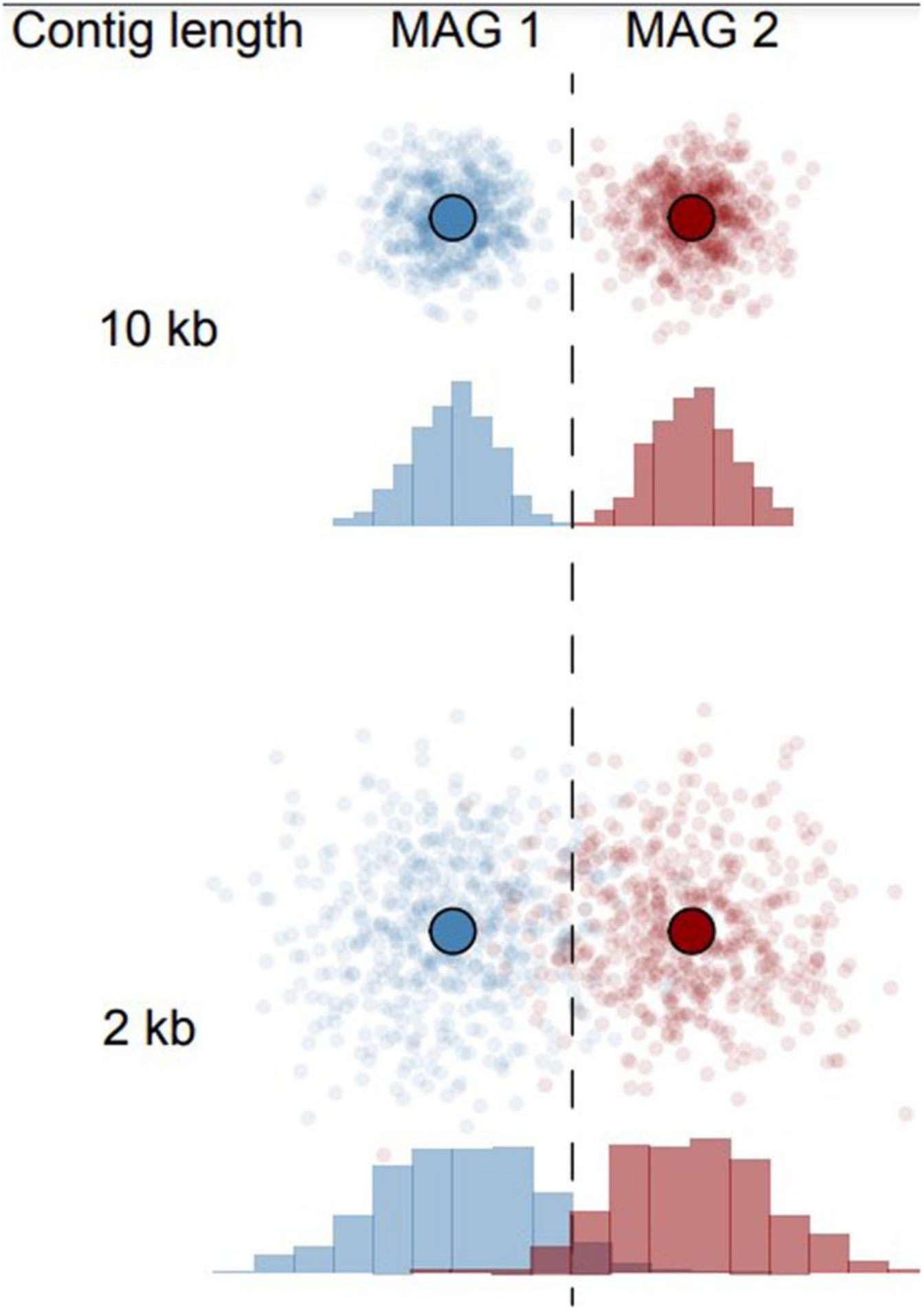


The x-axis represents the Assembly release date, ranging from 2000 to beyond 2020. The y-axis, labeled 'Number of prokaryotic genomes released at N C B I', ranges from 10 superscript 0 end superscript to 10 superscript 7 end superscript. Three lines are plotted: 1. Isolates plus M A G s (brown line) 2. Isolates with asterisk plus M A G s (light brown line) 3. M A G s (blue line). All three lines show an exponential growth trend, with the Isolates plus M A G s line consistently higher than the others. The M A G s line starts later, around 2010 and shows a sharp increase around 2020. The key observations in the graph are: In 2000, the number of Isolates plus M A G s was about 10.

By 2010, Isolates plus M A G s reached about 1000, while M A G s were just beginning to appear. Around 2020, there's a notable acceleration in growth for all categories, particularly M A G s. By the end of the graph, Isolates plus M A G s approach 10 superscript 7 end superscript, Isolates with asterisk plus M A G s and M A G s are slightly lower and reach about 10 superscript 6 end superscript.

Navigate back to Figure 9.1

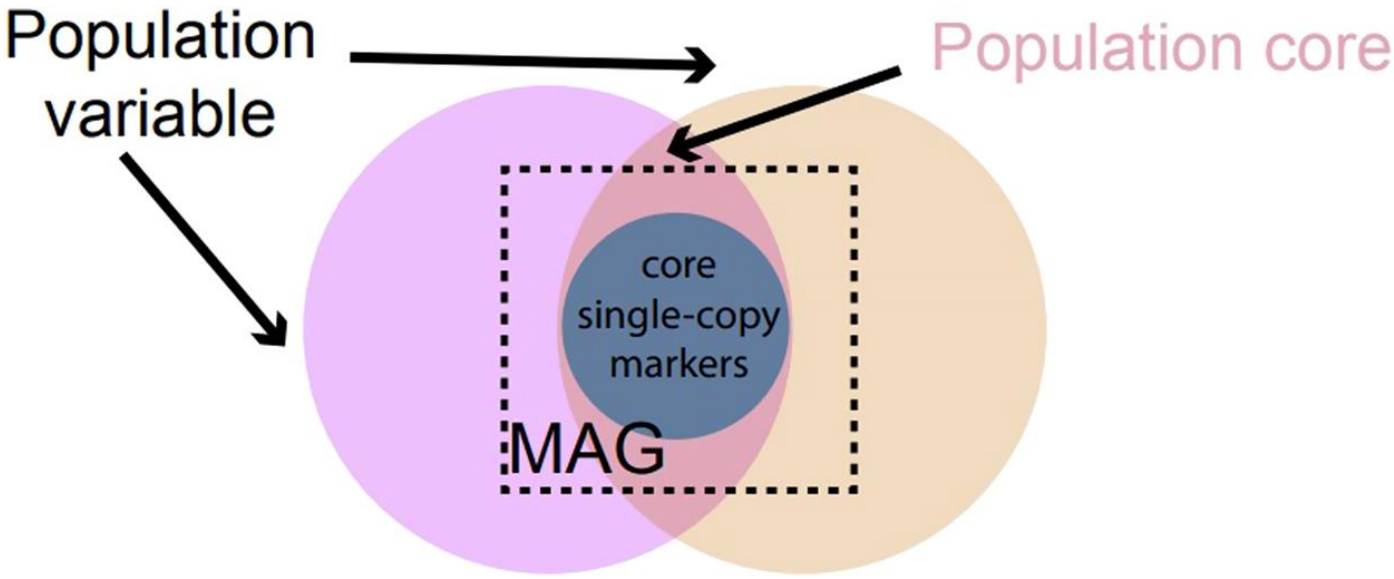
Description of Figure 9.2


Contig length
MAG 1
MAG 2
10 kb
2 kb

The image contains three column as follows: Contig length, M A G 1 and M A G 2. The contig lengths studied are 10 k b and 2 k b. For 10 k b, M A G 1 shows a blue circle with highly intense and dense scattered points around it followed by a bar graph which shows a symmetrical graph with peak near the center. For 10 k b, M A G 2 shows a red circle with highly intense and dense scattered points around it followed by a bar graph which shows a symmetrical graph with peak near the center. For 2 k b, M A G 1 shows a blue circle with a less intense and spread out scattered points around it followed by a bar graph similar to that of 10 k b M A G 1, but the bars are wider. For 2 k b, M A G 2 shows a red circle with a less intense and spread out scattered points around it followed by a bar graph similar to that of 10 k b M A G 2, but the bars are wider. For 2 k b, the points around the circles and the bar graphs overlap in the middle.

Navigate back to Figure 9.2

Description of Figure 9.3



The image shows a Venn diagram with two overlapping large circles both labeled 'Population variable'. The overlapping area contains a smaller circle labeled 'core single-copy markers' and it is enclosed inside a dashed square labeled 'Population core'. Also, the dashed square is labeled 'M A G' encompasses the overlapping area and the smaller circle.

Navigate back to Figure 9.3

CHAPTER 10

Understanding and modeling the structure of omics data through the lens of probability distributions

Kerry L. Ivey[1,2], [1]Division of Aging, Department of Medicine, Brigham and Women's Hospital, Boston, MA, United States, [2]Department of Medicine, Harvard Medical School, Boston, MA, United States

Abstract

Large amounts of information can be readily extracted from individual microbial community datasets. However the statistical distributions that best describe these data are not fully elucidated. This is a major problem as the distributional assumptions applied to statistical models of microbiome data affect many, if not all, stages of microbiomic inference. Laying the fundamental groundwork for our understanding of microbiome ecosystems, and how best to model them, this study aimed to determine the probability distributions and number of distribution components that best describe microbiome attributes in metagenomic datasets using the intelligent decision support framework (IDSF). By providing a detailed view of the distributional topography of the human stool microbiome this study brings us a step closer to understanding and modeling microbiome data in a biologically appropriate manner. It is apparent that, for metagenomic data, a one-size-fits-all approach to the underlying distributional assumptions of statistical analysis methodologies is not appropriate and that caution is needed when making

broad-based assumptions about the distribution of individual microbiome attributes.

Keywords

Metagenome; analysis; probability distribution; model; microbiome

Why we need new methodologies for determining probability distributions in microbiome studies

The human gastrointestinal microbiome is a major influence on disease development, progression, and response to therapy across a wide range of clinical contexts, particularly in metabolic and immune dysfunctions (Kau, 2011; Tilg & Kaser, 2011; Wiertsema, 2021). Recent advances in molecular biology have enabled, large population-based cohort, studies to assess the type and abundance of microbes that inhabit the gastrointestinal tract (by analyses based on the taxonomic composition inferred from sequencing marker genes), their genetic potential (by analyses based on whole metagenome sequencing), and their genetic activity (by analyses of metatranscriptome data) (Arumugam, 2011; Everett, 2021; Huttenhower, 2012). This new analytical capacity has allowed for a comprehensive investigation of human exposures, such as diet, lifestyle, and environment as they relate to the composition and behavior of complex resident microbial systems (Ivey, 2019; Merra, 2024; Mobegi, 2020; Spor et al., 2011). However the development and application of higher throughput assessment technologies have outpaced our understanding of the statistical and distributional properties of the data output from bioinformatic pipelines. This is of concern as a deep understanding of variable distributions, particularly dependent variable distributions, is the cornerstone of parametric statistics.

In the context of this chapter a microbiome attribute is a variable that is outputted from metagenomic bioinformatic pipelines in the relative abundance form. Examples include the relative abundance of different taxa

(at different taxonomic levels such as species or phyla) and the relative abundance of components of the metabolic processes Page 208 (from genes to pathways). There is an increased capacity to extract more information from single microbiome datasets, such as membership and functionality (composition and expression) (Beghini, 2021). As such it is more important than ever to revisit, at the host population level, the probability distribution of attributes comprising microbiome ecosystems, and to identify the factors that influence these distributions. Given that statistical analyses of microbiome-based hypotheses have the potential to influence our understanding of human health and its determinants (Ahn & Hayes, 2021; Cho & Blaser, 2012), we must be able to generate models that adequately reflect the observed data. The choice of the statistical model used to analyze data has special relevance to microbiome studies as it affects many, if not all, stages of microbiomic inference whereby conclusions are drawn, and predictions made, about the structure of the microbiome. The distribution of variables included in the analysis is a major determinant of the statistical model selection, particularly when considering the distribution of the dependent variable. It has been said that "understanding statistical distributions is a fundamental requisite to researchers in almost all disciplines" (Thomopoulos, 2017). As such it is necessary that we first define how microbiome data aredistributed across different microbiome attributes.

There are numerous approaches available to researchers conducting associative microbiome studies, each of which is based on different underlying statistical assumptions. Particularly, for parametric analyses, there is a lack of consensus regarding the probability distributions and number of distribution components that reflect the observed individual microbiome attributes (Mallick, 2017; Xia & Sun, 2017). It is clear that the distributions of attributes comprising a microbiome ecosystem, such as that of the human fecal microbiota, are not fully resolved. The existence of different distributional properties of microbiome attributes is a problem for the microbiome field, as it affects our capacity to validate findings across

studies. Furthermore, when different analytic methods are used in microbiome studies, metaanalysis of the results becomes extremely difficult because the estimates in each study are based upon different parameters. Therefore identifying the underlying distribution may help "standardize" the process of analysis and allow metalevel studies.

An attribute's distribution can be explained by one component (e.g., a single Gaussian curve for a normally distributed attribute) or more than one component, whereby the overall profile of an attribute's

distribution is comprised of several smaller, simpler distributions (e.g., a mix of two Gaussian curves representing two subgroups with different averages or variances). Probably the distribution of attributes at every level of taxonomic rank and genetic information has subcomponents of different prevalence in the host population. As such assumptions of one-component distributions may not always reflect the real observed data. Finite mixture modeling may be useful to determine the underlying probability distributions and number of components within a given host population (Lindsay, 1995; Melnykov & Maitra, 2010). This is because the theory of finite mixture models may allow us to get a statistical estimate of the number of components that best describe the data. This is all performed without the need to have an *a priori* knowledge of the distribution or the number of components, as is the case with microbiome data. Identification of the probability distribution that best describes the observed data may help in reducing bias and variance of the estimates in association studies.

As the availability of increasingly large microbiome sequence datasets presents ever more opportunities to understand the relationship of a broader set of attributes per sample it is more imperative than ever that we develop a detailed understanding of how these data are statistically distributed. Utilizing novel methodologies this study aimed to assess the distribution of individual microbiome attributes to identify determinants of distributional properties and to evaluate the implications for statistical analyses. To achieve this the intelligent decision support framework (IDSF) is presented. It is as

follows: A novel analytic pipeline for determining the probability distribution and number of distribution components that best describe the relative abundance of measured (observed) microbiome attributes whose distribution is not known *a priori*.

Men's lifestyle validation study is an excellent resource for studying distributional profiles of microbiome attributes

Cohort description

This study used data from the Men's Lifestyle Validation Study (MLVS) as a case example on which to implement the IDSF. The MLVS (Mehta et al., 2018) was a longitudinal sub-study nested within the Health Professionals Follow-up Study (HPFS) (Kasper et al., 2009) where US health professionals have been monitored from 1986 to present day. In 2012 a subset of MLVS participants aged 45 to 80 years, Page 210 who did not report the use of antibiotics within the past year, was recruited to provide fecal samples as part of the MLVS (n = 247 participants presented in this analysis). MLVS participants were free of coronary heart disease, stroke, cancer, or major neurological disease.

Metagenomic description of fecal samples

Using methods previously described (Abu-Ali et al., 2018; Mehta et al., 2018), over a period of 6 months, MLVS participants provided up to four fecal samples in accordance to the Study Protocol 22067–102 entitled "Men's Lifestyle Validation Study and Microbiome Correlation" which was approved by the Harvard T. H. Chan School of Public Health Institutional Review Board, and informed consent was obtained from all participants.

The processes for handling the fecal samples as well as the extraction, processing, and sequencing of the fecal DNA have been previously published (Abu-Ali et al., 2018; Mehta et al., 2018). The Nextera XT DNA Library

Preparation Kit for whole-genome sequencing was used. The Illumina HiSeq paired-end (2 × 101 nucleotides) sequencing platform was used to obtain whole metagenome shotgun sequence data. After completing the sequencing but before completing the biocomputing quality control the sequencing depth (mean±SD) for DNA was 3.8±1.6 giganucleotides (Gnt). After quality control the sequencing depth was 1.8±0.7 Gnt. Human sequences were then removed as well as quality trimmed and depleted duplicate reads using HMP protocols with KneadData (McIver et al., 2018). Taxonomic profiling was performed using MetaPhlAn2 (Truong, 2015), and functional profiling of genes was performed using HUMAnN2 (Franzosa, 2018).

Intelligent decision support framework methodology

In the pursuit of model precision the IDSF aims to: (1) identify the probability distribution and number of distribution components that best describe each individual observed microbiome attribute and (2) apply finite mixture modeling to identify the parameters of the model that best reflect a group of observed attributes. **Fig. 10.1** outlines the four components of the pipeline: Component 1 where the probability distribution and the number of distribution components of each microbiome attribute are identified; Components 2 and 3 where we identify the level of precision and optimization method that we should be used for all analyses within a given attribute group; and Component 4, where the Page 211 microbiome attribute (dependent variable) is regressed against an exposure (independent) variable, such as consumption of sugar-sweetened beverages, in a variety of adjusted settings (Ivey et al., 2018).

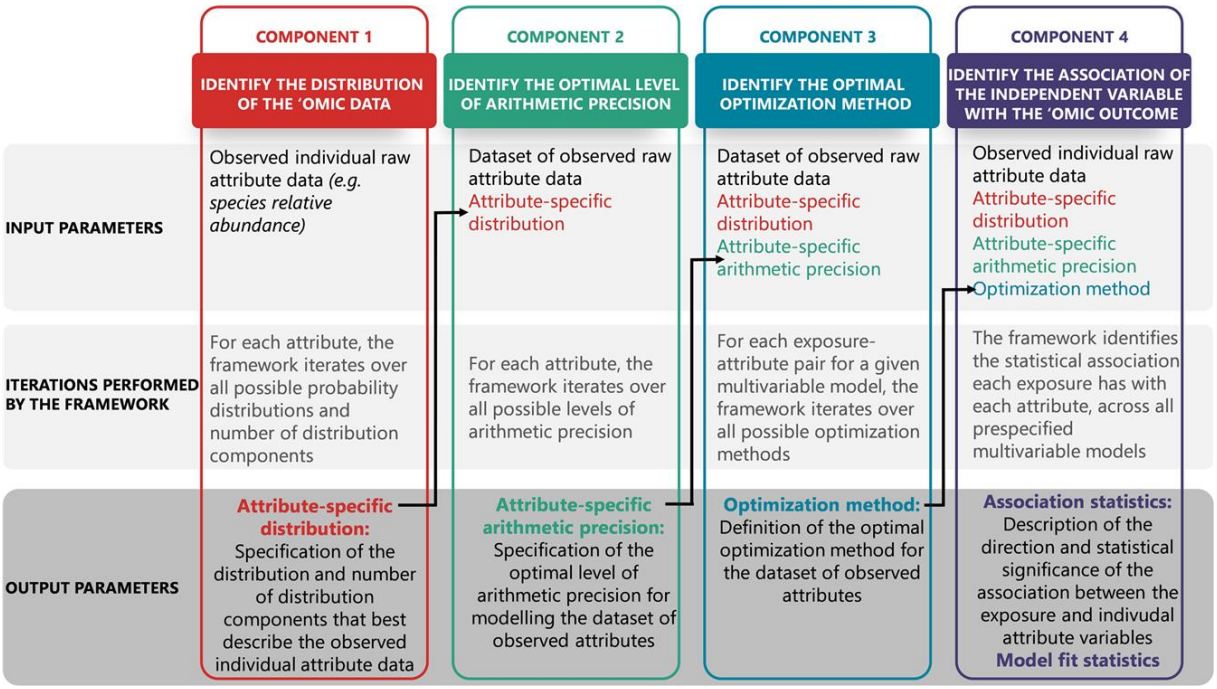

COMPONENT 1
IDENTIFY THE DISTRIBUTION OF THE 'OMIC DATA
COMPONENT 2
IDENTIFY THE OPTIMAL LEVEL OF ARITHMETIC PRECISION
COMPONENT 3
IDENTIFY THE OPTIMAL OPTIMIZATION METHOD
COMPONENT 4
IDENTIFY THE ASSOCIATION OF THE INDEPENDENT VARIABLE WITH THE 'OMIC OUTCOME
INPUT PARAMETERS
Observed individual raw attribute data (e.g. species relative abundance)
Dataset of observed raw attribute data
Attribute-specific distribution
Dataset of observed raw attribute data
Attribute-specific distribution
Attribute-specific arithmetic precision
Observed individual raw attribute data
Attribute-specific distribution
Attribute-specific arithmetic precision
Optimization method
ITERATIONS PERFORMED BY THE FRAMEWORK
For each attribute, the framework iterates over all possible probability distributions and number of distribution components
For each attribute, the framework iterates over all possible levels of arithmetic precision
For each exposure-attribute pair for a given multivariable model, the framework iterates over all possible optimization methods
The framework identifies the statistical association each exposure has with each attribute, across all prespecified multivariable models
OUTPUT PARAMETERS
Attribute-specific distribution: Specification of the distribution and number of distribution components that best describe the observed individual attribute data
Attribute-specific arithmetic precision: Specification of the optimal level of arithmetic precision for modelling the dataset of observed attributes
Optimization method: Definition of the optimal optimization method for the dataset of observed attributes
Association statistics: Description of the direction and statistical significance of the association between the exposure and indivudal attribute variables
Model fit statistics

FIGURE 10.1 Overview of the components comprising the intelligent decision support framework. Adapted from Ivey et al. (2018).

Follow for extended description

Component 1 of the intelligent decision support framework

Component 1 is referred to as the "intelligent" aspect of the framework as, for each microbiome attribute, it identifies and selects both the number of distribution components and the type of probability distribution that best describes the observed data. To estimate parameters and compare the capacity of models to reflect observed data, the framework iterates over possible log-likelihood distributions for the dependent variable, which include the normal (Gaussian), truncated-normal, folded-normal, inverse Gaussian, beta, gamma, log-normal, truncated log-normal, Student's *t*, and Weibull probability distributions (Thomopoulos, 2018).

Methodologic rationale

Because the distributional profiles of microbiome data are not clearly defined, we did not know *a priori* what the precise distribution and number of distribution components were for each microbiome attribute. As such we aimed to identify the probability distribution and number of distribution components that best describe each individual Page 212 observed microbiome attribute, in its raw, untransformed state. This was achieved through the implementation of finite mixture modeling (***SAS/STAT® 13.1 User's Guide*, 2013**), which is ideally suited for addressing this aim as it is capable of identifying distributions in the heterogeneous data of mixed distributions. When exploring mixtures finite mixture models are capable of providing two pieces of information: mixture probabilities (π_j) and component distributions (p_j). As such the general expression for the finite mixture model included in the framework was as follows: $f(y) = \Sigma_{j=1}^{k} \pi_j p_j(y; x'_j \beta_j, \varnothing_j)$. Mixture probabilities are only computed if a microbiome attribute has more than one component in the distribution (k). The values of p depend upon the regressor value x_j, the regression parameter β_j, and, in some cases, the scale parameter $\varnothing_j$. The total model component distributions reflect the sum of p starting at $j = 1$, the first component, and ending with $j = k$, the maximum number of components in the mixture.

Using the high-performance finite mixture modeling SAS procedure (***SAS/STAT® 14.1 User's Guide*, 2015**) and a custom SAS macro based on the iterator design pattern, for each microbiome attribute in a given dataset, the type of probability distribution, and number of distribution components that best describe the observed distribution of that attribute was identified. To do this the distribution-specific model was identified by iterating over the possible log-likelihood distributions that relative abundance data could take and then identifying the number of distribution components that best described the observed data, in the most parsimonious fashion. This is achieved by selecting the number of components that produce the smallest Bayesian information criterion (BIC), which penalizes for model complexity. As such there are two processes at play: (1) identifying the "distribution-

specific models" that reflect the optimal number of distribution components for each unique distribution and (2) selecting the single "distribution-specific model" that best reflects the observed data.

Identifying the "distribution-specific models" that reflect the optimal number of distribution components for each unique probability distribution

For every individual microbiome attribute and each of the assessed distributions, there will be an optimal number of distribution components that best describe the observed data. Therefore this approach iterates over all microbiome attributes, and then for each attribute, iterates over all distributions to find, for each distribution, the most Page 213 parsimonious distribution-specific model that best describes the observed attribute data. Although the pipeline can be modified to include any number of possible probability distributions, to ensure models are not overly complex and cumbersome to interpret, in this study, the number of possible distribution components explored was either one, two, or three components. Therefore, to determine the number of components that contribute to the overall distribution of a particular microbiome attribute, for each probability distribution, the pipeline repeats the finite mixture model three times, once for each possible number of components. The BIC which penalizes model complexity is outputted. By selecting the number of components (either one, two, or three) that produce the smallest BIC for a given distribution, the finite mixture model identifies the most parsimonious "distribution-specific model" that best describes the observed attribute data in a given dataset.

Identifying the "attribute-specific model" by selecting the single "distribution-specific model" that best reflects the observed data

The suitability of these "distribution-specific models" was then compared by ranking, in ascending order, the Akaike's information criterion (AIC), ias-

corrected AIC, BIC, and deviation of Pearson Statistic from sample size. By selecting the distribution with the lowest sum of these four ranks, the distribution, and henceforth number of distribution components, that best described the observed distribution was identified. This selected distribution is specific for a given microbiome attribute and as such represents the "attribute-specific model" for that specific data. Sensitivity analyses including different combinations of the statistics yielded similar results.

Components 2 and 3 of the intelligent decision support framework

Components 2 and 3 represent the "decision support" aspect of the framework. This is because both components provide feedback statistics to the investigator for manual determination of the numerical precision and optimization method to be applied to an entire group of microbiome attributes in an analytic dataset.

Arithmetic precision

This framework determines the level of arithmetic precision that brings forth the model that best represents the observed data. Specifically, for Page 214 each attribute-specific model, the algorith repeats the finite mixture model for all values of arithmetic precision, ranging from 10^{-3} to the maximum number of decimal places outputted by the bioinformatic pipeline used to generate the data. The range-standardized residuals of all microbiome attributes are compared across all levels of arithmetic precision, and summary statistics are outputted for the investigator to determine the level of arithmetic precision to be used in subsequent components of the framework.

Suboptimal definitions of precision can lead to misleading results in association studies, primarily because of the effect of arithmetic precision on errors due to rounding and on our capacity to observe subtle variations in the raw data. By strategically evaluating the optimal level of arithmetic precision

this approach minimizes computational artifacts and balances the prevention of both overfitting to noise and any loss of information.

Optimization

Incorporating the attribute-specific model and arithmetic precision Component 3 implements multivariate-adjusted exposure-attribute finite mixture models, iterating over all possible optimization methods, including conjugate-gradient optimization; double-dogleg optimization; Newton-Raphson optimization with ridging; Newton-Raphson optimization with ridging and a line-search algorithm; Nelder-Mead simplex optimization; Dual Quasi-Newton optimization; trust-region optimization; and no optimization. The range-standardized residuals and optimization statistics of all microbiome attributes are compared across all optimization methods, and summary statistics are outputted for the investigator to make study-level determinations as to the optimization method to be implemented in the final component.

Particularly in highly dimensional data, the choice of optimization method can substantially impact observed associations and even the likelihood that the model converges. Importantly the choice of optimization methods impacts the fit and residuals of the model, thus dictating how well the model we are building reflects the real, observed relationships.

Component 4 of the intelligent decision support framework

Component 4 represents the “precision” aspect of the framework whereby all decisions made in the preceding components feed into this component to develop an exposure-attribute statistical model that Page 215 incorporates the most appropriate probability distribution and distribution components. The model is implemented in both unadjusted and adjusted settings using a level of arithmetic precision that reduces noise due to unnecessary measurement error while simultaneously enhancing biological

representation whereby the identified associations are more likely to reflect the relationships that underlie the biological state of the microbiome. Furthermore, by incorporating the most appropriate optimization method, the framework ensures that the "precision model" output is comprised of the most optimal estimates for a given microbiome attribute group.

The final "intelligent decision" in the pipeline relates to the assessment of the minimum number of nonzero values of an attribute in the sample from which a biologically representative continuous model can be derived. Varying distributions of the exposure variable and nonzero microbiome attributes, as well as varying sample size, mean that the value of this threshold can vary across exposures, attributes, attribute groups, and studies. As such the inclusion threshold for the prevalence of nonzero values cannot be determined *a priori* and is instead determined, at an attribute-specific level, in Component 4 of the pipeline.

If the range-standardized residual of the resultant exposure-attribute model is less than a user-specified threshold, the framework will output, for each exposure-attribute-adjustment combination, the following: (1) fit statistics describing how well the model describes the observed data, (2) coefficients describing how a given exposure is associated with the attribute, (3) standard error and confidence intervals provide information on the precision and reliability of the estimates, and (4) associated *P* values for statistical significance. If the number of participants with nonzero values is insufficient and the range-standardized residuals exceed a user-specified threshold, the finite mixture model analysis is repeated with the microbiome attribute represented as a binary (detected vs not detected) variable.

For analyses with continuous dependent variables calculations of the minimum sample size required to adequately test the hypothesis depend upon the accepted risk of Type I and Type II statistical errors. Given that sample size calculations are highly experiment-specific and reliant on decisions by the investigator, the IDSF aims to make minimal microbiome attribute exclusions. Given that there is no clear guidance in the literature for

a minimal sample size to assess distributions, the IDSF enables researchers to dictate the size of the final dataset entered into the framework. Page 216

Rationale for the intelligent decision support framework

The rationale for the “intelligent” and “decision support” distinction lies in the number of potential decisions to be made and the requirements for cross-validation. Regarding the “intelligent” componentprobability distributions and the number of distribution components are determined individually for each microbiome attribute. Given that the number of attributes in the attribute group can range into the thousands, it is unfeasible for Component 1 to incorporate investigator feedback. Given that the distribution of microbiome attributes can vary across host populations, all attribute-specific distributions are determined by the framework on an individual host population level.

Conversely the choice of arithmetic precision and optimization method is made at the level of attribute groups to allow comparison of coefficients for attributes within an attribute group. Although the decisions in Components 2 and 3 could be automated, external factors, such as an investigator’s desire to compare regression coefficients across different studies, may also play a role in determining what level of arithmetic precision and optimization method should be used. As such the framework feeds back summary statistics to “support” the decision of the investigator.

Frequently used in ecological analyses finite mixture modeling allows for the classification of individuals based on unobserved heterogeneity, brought forth by unobserved subpopulations in the sample. Not only does this allow us to model the sample as a whole but also allows us to explore distinct associations within each subpopulation as well as the contribution to component membership. As such this decision support framework, for each individual microbiome attribute, identifies the unique probability distribution and distribution components that best describe the observed data. By

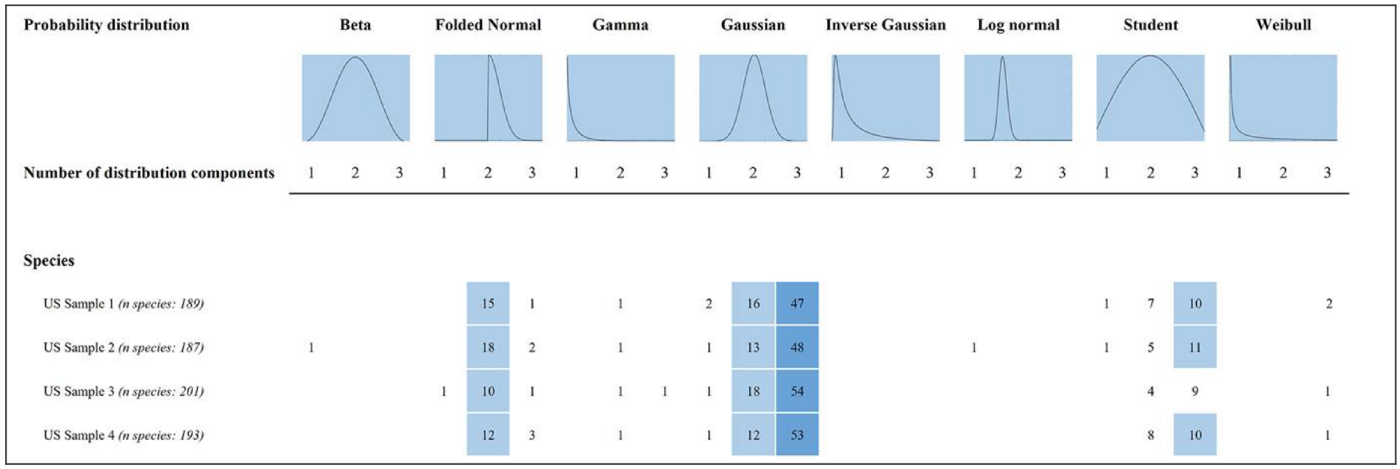
improving our capacity to build statistical models that adequately reflect the observed data of host populations this framework has the potential to facilitate precision in finite mixture modeling and achieve superior model fit than traditional regression-based analyses.

Finite mixture modeling represents a method for describing heterogeneity in host population distributions. However the most appropriate method for selecting the optimal number of distribution components is not fully resolved, and more work in this area is also warranted. To overcome this challenge multiple model fit metrics to identify the optimal number of distribution components were simultaneously used. Page 217 Another limitation of finite mixture modeling is that it is unknown whether the identified number of distribution components truly reflects the heterogenous distribution in the observed data. To address this issue multiple sample collection time points were incorporated into the analysis. From this application of the intelligent decision framework to samples collected at different time points, it was observed that species data have a pattern of distribution that is overrepresented in the two-component folded normal, and two- and three-component Gaussian distributions. This overrepresentation could be the result of a microbiome ecosystem composed of multiple species with very similar properties, roles, or functions, such as those that occur when different organisms fill the same ecological niche. However the implications of this overrepresentation of different types of distributions are not fully understood, and further research in this area is indicated. This analysis assumes homogenous mixtures of distribution components. Although there is no reason to assume that this assumption would be violated in real datasets, it is possible that the true distribution mixtures may comprise more than one distribution type. However given that introducing different distributions for each component will introduce unnecessary complexity at the cost of usefulness, a homogenous mixture is an appropriate assumption for this type of data. However there is a need for further studies in this area to confirm that this homogeneous approach is ideal.

Observed distributions of microbiome attributes as identified using the intelligent decision support framework

Implementation of Component 1 of the IDSF (**Fig. 10.1**) allowed us to observe what continuous probability distribution best represents the relative abundance of attributes that comprise the human fecal microbiome. No one distribution described all attributes within an attribute type, but rather, datasets of each attribute type comprised a mixture of distributions. One attribute observed across the majority of attribute types was the overrepresentation by the three-component Gaussian distribution in the distributional profiles. **Table 10.1** summarizes previously published results (Ivey et al., 2018) from four fecal collections that demonstrate that at the species taxonomic level the most frequently observed distributions were two-component folded normal and the two- and three-component Gaussian distributions. Page 218

Table 10.1



Probability distribution	Beta			Folded Normal			Gamma			Gaussian			Inverse Gaussian			Log normal			Student			Weibull		
Number of distribution components	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Species																								
US Sample 1 *(n species: 189)*					15	1		1			2	16	47						1	7	10			2
US Sample 2 *(n species: 187)*	1				18	2		1			1	13	48			1			1	5	11			
US Sample 3 *(n species: 201)*				1	10	1		1		1	1	18	54							4	9			1
US Sample 4 *(n species: 193)*					12	3		1			1	12	53							8	10			1

Results are presented as percentage of attributes meeting criteria. Colored highlighting identifies frequently occurring distributions (≥10% of attributes), with darker colors representing the most frequently identified distribution–component pair for a given sample set.

Adapted from Ivey et al. (2018).

The specific determinants of the distribution of an attribute need to be explored. The percentage of relative abundances with a value of zero and the mean attribute relative abundance when including both zero and nonzero values as well as when only considering nonzero values should also be explored for their role in influencing distributional profiles. As a next step it is extremely important that the contributors to an attribute's distribution be defined especially when conducting analyses that assume one distribution type for all microbiome attributes, as it has the potential to disproportionally and nonrandomly bias the results of the analyses based on the distribution determinants. The nonrandom nature of an attribute's distribution further highlights the need to conduct analyses that are based on appropriate distributional assumptions.

Irrespective of what is causing these distributional profiles, it is evident that, given the vast array of distributions within a given taxonomic rank, a one-size-fits-all approach is not appropriate for microbiome data. This is why

a framework capable of modeling data based on its given distributional properties is indicated.

Application of the intelligent decision support framework to assess diet-microbiome associations

The IDSF has been applied to two distinct cohorts to unravel the specific associations between the consumption of sugar-sweetened beverages and gut microbiome composition. Using previously published results Page 219 (Ivey et al., 2018), even after adjusting for potential confounders (including geographical location, alcohol intake, gender, hypertension, smoking status, physical activity, and potential dietary confounders) and correction for multiple hypothesis testing (Bonferroni), numerous significant relations of sugar-sweetened beverage intake were identified. These included relative abundances of the genus *Alistipes* and the species *Eubacterium ramulus* and *Enterobacter cloacae* (Ivey et al., 2018). It is interesting to note that these associations were not detected with the use of traditional regression-based methodologies that are often applied to microbiome data. Using a two-sample *t*-test to compare the standardized residuals resulting from the IDSF were smaller than those produced by other regression methods for 100% of the species. This indicates that the models generated as part of the IDSF fitted the observed species relative abundance data well, with the least amount of error.

The intelligent decision support framework in context

Tools for the statistical analysis of microbiome data rely on a variety of assumptions for the underlying distribution of the data, as well as statistical normalization techniques (Cao, 2016; Mallick, 2017; Xia & Sun, 2017). This issue of data normalization and probability distribution of the data has

prompted consideration of which normalization functions are most appropriate. Current approaches to parametric analysis of microbiome relative abundance data are based on the assumption that the distribution of microbiome attributes follows either an unimodal or a multimodal distribution (Cao, 2016; Mallick, 2017; Paulson et al., 2013; Xia & Sun, 2017). For unimodal analyses current dogma suggests that microbiome attribute data are not normally distributed but can be normalized through the application of statistical normalization functions. This in-depth analysis of the probability distributions that make up human fecal microbiome ecosystems demonstrates that distributional profiles of microbiome attributes vary within and across attribute types, and as such there is no single method that would be capable of transforming all microbiome attributes into the one component of Gaussian distributions that are often assumed to be present in many statistical analysis protocols for microbiome data. This suggests that for Microbiome Wide Association Studies that assume unimodal distributions with fixed distribution across attributes, statistical normalization may not be the most appropriate approach as it does not universally transform data into a synchronously distributed dataset. Page 220

A counterproposal to the assumption that all data are distributed unimodally is that individual microbiome attributes may be a composite of more than one distribution component, as is evidenced by the frequent use of zero-inflated modeling techniques. One example of this multimodal assumption is the zero-inflated beta regression model with random effects pipeline (Chen & Li, 2016) which assumes that the raw, untransformed total-sum scaled data have a zero-inflated beta distribution. Although not ubiquitous, irrespective of data source or attribute type, the most commonly occurring distribution identified was the three-component Gaussian distribution.

The development of tools with various distributional assumptions (Mallick, 2017; Xia & Sun, 2017) suggests there is increased recognition as to the

importance of data distributions. Given that how data are distributed impacts the type of statistical tests performed, and also how we interpret results from statistical analyses, an evaluation of data distribution from taxonomic, DNA-seq, and RNA-seq data from the same sample is critical. Particularly for species data the observed probability distribution of microbiome attributes follows a predictable and reproducible pattern that is unique to a given attribute type. The vast array of distributions within a given taxonomic rank and the changing distributional characteristics across varying levels of taxonomy highlight why a one-size-fits-all approach is not ideal for microbiome data, and why a framework capable of modeling data based on its given distributional properties is indicated. The differing effects of normalization within and between taxonomic ranks add further support for the need for an analytic methodology that addresses the unique distributional properties of each microbiome variable. Microbiome data have a unique combination of distributional attributes, which often include zero inflation and multimodal distributions. Given this complexity a deep understanding of attribute distributions is particularly important for microbiome analyses, as the outcomes of these analyses have the potential to have a large impact on our understanding of the biological rules influencing the structure of microbiomes. By being able to model the data according to the individual distribution parameters of each attribute metagenomic studies will be better equipped to build models that adequately reflect the observed data, therefore, increasing their chances of unveiling biologically meaningful associations.

This case study focusing on stool microbiome demonstrates that care must be taken when making broad-based assumptions about the distribution of individual microbiome attributes that make up a dataset. Page 221 We are now in the era where the sizes of host population studies are getting larger and larger where co-analysis of multiple body sites (such as analyzing oral and gut, or vagina and stool microbiomes together) where attributes might have different distributions is becoming more prevalent (Everett, 2021; Keshet & Segal, 2024; Kunath et al., 2024). With the microbiome being

increasingly recognized as an important determinant of human health, it is more important than ever to build biologically relevant models of microbiome data. As the distributional assumptions of a model are a major determinant of the ability of a model to reflect observed data, by providing a detailed view of the distributional topography of the human stool microbiome, this study has brought us one step closer toward understanding and modeling microbiome data in a biologically appropriate manner. Further replication is needed to confirm the biological associations we observed extend beyond the host population reported here, and simulation studies will be useful in determining the performance of the IDSF.

Acknowledgments

Gratitude is extended to the co-authors of the original abstract, which include Andrew T Chan, Geraint Rogers, Ruifeng Li, Ambika Sajita, Peter Kraft, Robyn McDermott, Steve Wesselingh, G Abu-Ali, H Malik, Walter Willett, Curtis Huttenhower, Jacques Izard, and Eric B Rimm. Their invaluable discussions and contributions greatly enriched this work. Additionally deep gratitude is extended to the participants of the MLVS and HPFS cohorts, whose commitment and dedication to advancing population sciences have been instrumental in enabling research that aims to improve the health and well-being of individuals and communities worldwide.

References

Abu-Ali et al., 2018 Abu-Ali GS, Mehta RS, Lloyd-Price J, et al. Metatranscriptome of human faecal microbial communities in a cohort of adult men. *Nature Microbiology*. 2018;3(3):356–366 **https://doi.org/10.1038/s41564-017-0084-4**.

Ahn and Hayes, 2021 Ahn J, Hayes RB. Environmental influences on the human microbiome and implications for noncommunicable disease. *Annual Review of Public Health*. 2021;42(1):277–292

https://doi.org/10.1146/annurev-publhealth-012420-105020.

Arumugam et al., 2011 Arumugam M, Raes J, Pelletier E, et al. Enterotypes of the human gut microbiome. *Nature*. 2011;473(7346):174–180 **https://doi.org/10.1038/nature09944**.

Beghini et al., 2021 Beghini F, McIver LJ, Blanco-Míguez A, et al. *Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3*. *Elife*. 2021;10 e65088 **https://doi.org/10.7554/eLife.65088**.

Cao et al., 2016 Cao K-AL, Costello ME, Lakis VA, et al. *MixMC: A multivariate statistical framework to gain insight into microbial communities*. *PLoS One*. 2016;11 e0160169 **https://doi.org/10.1371/journal.pone.0160169**.

Chen and Li, 2016 Chen EZ, Li H. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics (Oxford, England)*. 2016;32(17):2611–2617 **https://doi.org/10.1093/bioinformatics/btw308**.

Cho and Blaser, 2012 Cho I, Blaser MJ. The human microbiome: At the interface of health and disease. *Nature Reviews Genetics*. 2012;13(4):260–270 **https://doi.org/10.1038/nrg3182**.

Everett et al., 2021 Everett C, Li C, Wilkinson JE, et al. Overview of the microbiome among nurses study (Micro-N) as an example of prospective characterization of the microbiome within cohort studies. *Nature Protocols*. 2021;16(6):2724–2731 **https://doi.org/10.1038/s41596-021-00519-z**.

Franzosa et al., 2018 Franzosa EA, Mclver LJ, Rahnavard G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods*. 2018;15(11):962–968 **https://doi.org/10.1038/s41592-018-0176-y**.

Huttenhower et al., 2012 Huttenhower C. Structure, function and diversity of the healthy human microbiome. *Nature*. 2012;486(7402):207 **https://doi.org/10.1038/nature11234**.

Ivey, 2018 Ivey KL, Chan A, Rogers G, et al. Sugar-sweetened beverages and the microbiome: Implementation of intelligent decision support for precision in statistical modelling. *Current Developments in Nutrition*. 2018;2(11):nzy045 **https://doi.org/10.1093/cdn/nzy045**.

Ivey et al., 2019 Ivey K, Chan AT, Izard J, Cassidy A, Rogers GB, Rimm EB. Role of dietary flavonoid compounds in driving patterns of microbial community assembly. *mBio*. 2019;10:e01205–e01219 **https://doi.org/10.1128/mBio.01205-19**.

Kasper et al., 2009 Kasper JS, Liu Y, Giovannucci E. Diabetes mellitus and risk of prostate cancer in the health professionals follow-up study. *International Journal of Cancer*. 2009;124(6):1398–1403 **https://doi.org/10.1002/ijc.24044**.

Kau et al., 2011 Kau AL, Ahern PP, Griffin NW, Goodman AL, Gordon JI. Human nutrition, the gut microbiome and the immune system. *Nature*. 2011;474(7351):327–336 **https://doi.org/10.1038/nature10213**.

Keshet and Segal, 2024 Keshet A, Segal E. Identification of gut microbiome features associated with host metabolic health in a large population-based cohort. *Nature Communications*. 2024;15 **https://doi.org/10.1038/s41467-024-53832-y**.

Kunath et al., 2024 Kunath BJ, De Rudder C, Laczny CC, et al. The oral–gut microbiome axis in health and disease. *Nature Reviews Microbiology*. 2024;22(12):791–805 **https://doi.org/10.1038/s41579-024-01075-5**.

Lindsay, 1995 Lindsay BG. *Mixture models: Theory, geometry and applications* Institute of Mathematical Statistics 1995; **https://doi.org/10.1214/cbms/1462106013**.

Mallick et al., 2017 Mallick H, Ma S, Franzosa EA, Vatanen T, Morgan XC, Huttenhower C. Experimental design and quantitative analysis of microbial community multiomics. *Genome Biology*. 2017;18(1):228 **https://doi.org/10.1186/s13059-017-1359-z**.

McIver et al., 2018 McIver LJ, Abu-Ali G, Franzosa EA, et al. BioBakery: A meta'omic analysis environment. *Bioinformatics (Oxford, England)*. 2018;34(7):1235–1237 **https://doi.org/10.1093/bioinformatics/btx754**.

Mehta et al., 2018 Mehta RS, Abu-Ali GS, Drew DA, et al. Stability of the human faecal microbiome in a cohort of adult men. *Nature Microbiology*. 2018;3(3):347–355 **https://doi.org/10.1038/s41564-017-0096-0**.

Melnykov and Maitra, 2010 Melnykov V, Maitra R. Finite mixture models and model-based clustering. *Statistics Surveys*. 2010;4:80–116 **https://doi.org/10.1214/09-SS053**.

Merra et al., 2024 Merra G, Gualtieri P, La Placa G, et al. The relationship between exposome and microbiome. *Microorganisms*. 2024;12(7):1386 **https://doi.org/10.3390/microorganisms12071386**.

Mobegi et al., 2020 Mobegi FM, Leong LE, Thompson F, et al. *Intestinal microbiology shapes population health impacts of diet and lifestyle risk exposures in Torres Strait Islander communities*. *Elife*. 2020;9 e58407 **https://doi.org/10.7554/elife.58407**.

Paulson et al., 2013 Paulson JN, Olson ND, Braccia DJ, et al. metagenomeSeq: Statistical analysis for sparse high-throughput sequencing. *Bioconductor package* 2013; **http://www.cbcb.umd.edu/software/metagenomeSeq**.

SAS/STAT® 13.1 User's Guide, 2013 SAS/STAT® 13.1 User's Guide. (2013). SAS Institute Inc, SAS Institute Inc.

SAS/STAT® 14.1 User's Guide, 2015 SAS/STAT® 14.1 User's Guide. (2015). SAS Institute Inc, SAS Institute Inc.

Spor et al., 2011 Spor A, Koren O, Ley R. Unravelling the effects of the environment and host genotype on the gut microbiome. *Nature Reviews Microbiology*. 2011;9(4):279 **https://doi.org/10.1038/nrmicro2540**.

Thomopoulos, 2017 Thomopoulos NT. *Statistical distributions: Applications and parameter estimates* Cham, Switzerland: Springer; 2017;.

Thomopoulos, 2018 Thomopoulos NT. *Probability distributions* Cham, Switzerland: Springer; 2018;.

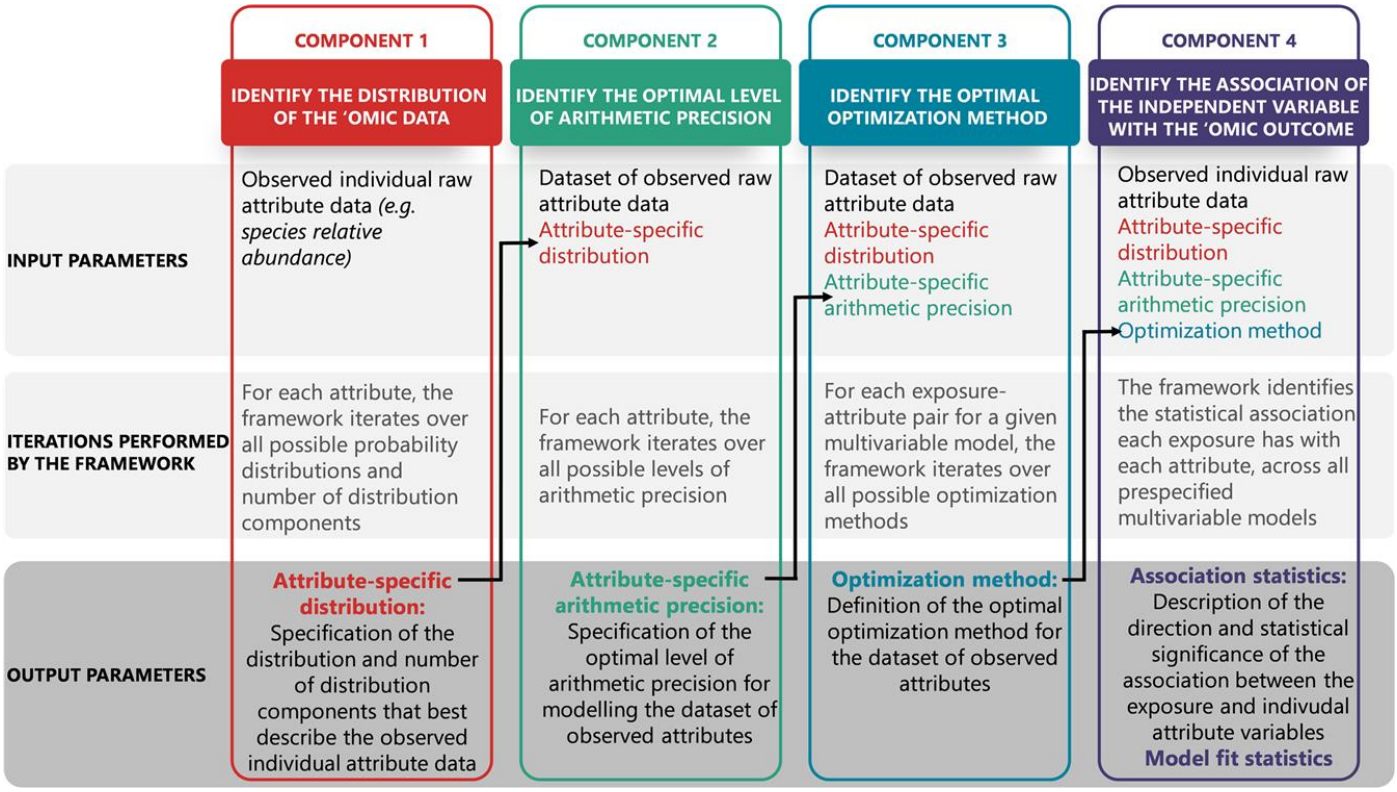
Tilg and Kaser, 2011 Tilg H, Kaser A. Gut microbiome, obesity, and metabolic dysfunction. *Journal of Clinical Investigation*. 2011;121(6):2126–2132 **https://doi.org/10.1172/JCI58109**.

Truong et al., 2015 Truong DT, Franzosa EA, Tickle TL, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*. 2015;12(10):902 **https://doi.org/10.1038/nmeth.3589**.

Wiertsema et al., 2021 Wiertsema SP, van Bergenhenegouwen J, Garssen J, Knippels LMJ. The interplay between the gut microbiome and the immune system in the context of infectious diseases throughout life and the role of nutrition in optimizing treatment strategies. *Nutrients*. 2021;13(3):886 **https://doi.org/10.3390/nu13030886**.

Xia and Sun, 2017 Xia Y, Sun J. Hypothesis testing and statistical analysis of microbiome. *Genes and Diseases*. 2017;4(3):138–148 **https://doi.org/10.1016/j.gendis.2017.06.001**.

Description of Figure 10.1


COMPONENT 1
IDENTIFY THE DISTRIBUTION OF THE 'OMIC DATA
COMPONENT 2
IDENTIFY THE OPTIMAL LEVEL OF ARITHMETIC PRECISION
COMPONENT 3
IDENTIFY THE OPTIMAL OPTIMIZATION METHOD
COMPONENT 4
IDENTIFY THE ASSOCIATION OF THE INDEPENDENT VARIABLE WITH THE 'OMIC OUTCOME
INPUT PARAMETERS
Observed individual raw attribute data (e.g. species relative abundance)
Dataset of observed raw attribute data
Attribute-specific distribution
Dataset of observed raw attribute data
Attribute-specific distribution
Attribute-specific arithmetic precision
Observed individual raw attribute data
Attribute-specific distribution
Attribute-specific arithmetic precision
Optimization method
ITERATIONS PERFORMED BY THE FRAMEWORK
For each attribute, the framework iterates over all possible probability distributions and number of distribution components
For each attribute, the framework iterates over all possible levels of arithmetic precision
For each exposure-attribute pair for a given multivariable model, the framework iterates over all possible optimization methods
The framework identifies the statistical association each exposure has with each attribute, across all prespecified multivariable models
OUTPUT PARAMETERS
Attribute-specific distribution: Specification of the distribution and number of distribution components that best describe the observed individual attribute data
Attribute-specific arithmetic precision: Specification of the optimal level of arithmetic precision for modelling the dataset of observed attributes
Optimization method: Definition of the optimal optimization method for the dataset of observed attributes
Association statistics: Description of the direction and statistical significance of the association between the exposure and indivudal attribute variables
Model fit statistics

The image contains three rows labeled as input parameters, interations performed and output parameters. Also, presents four components as follows: Component 1 is titled 'Identify the Distribution of the 'Omic Data' and involves observed individual raw attribute data (example: species relative abundance) as input parameters. The framework iterates over all possible probability distributions and number of distribution components. The output parameter is the attribute-specific distribution, specifying the distribution and number of distribution components that best describe the observed individual attribute data. Component 2 is titled 'Identify the Optimal Level of Arithmetic Precision' and involves a dataset of observed raw attribute data with attribute-specific distribution as input parameters. The framework iterates over all possible levels of arithmetic precision. The output parameter is the attribute-specific arithmetic precision, specifying the optimal level of arithmetic precision for modeling the dataset of observed attributes.

Component 3 is titled 'Identify the Optimal Optimization Method' and involves a dataset of observed raw attribute data with attribute-specific distribution and arithmetic precision as input parameters. For each exposure-attribute pair for a given multivariable model, the framework iterates over all possible optimization methods. The output parameter is the optimization method, defining the optimal optimization method for the dataset of observed attributes. Component 4 is titled 'Identify the Association of the Independent Variable with the 'Omic Outcome' and involves observed individual raw attribute data with attribute-specific distribution, arithmetic precision and optimization method as input parameters. The framework identifies the statistical association each exposure has with each attribute, across all prespecified multivariable models. The output parameters are association statistics, describing the direction and statistical significance of the association between the exposure and individual attribute variables and model fit statistics. The output parameters from one components leads to the input parameters in the next component.

Navigate back to Figure 10.1

CHAPTER 11

Metabolomics in microbial ecology: sample preservation, storage, and integration with metagenomics

Brenan Durainayagam[1], Kelli D. Goodman[2], Adam D. Kennedy[2] and Annie M. Evans[2], [1]Metabolon, London, United Kingdom, [2]Metabolon, Morrisville, NC, United States

Abstract

As technologies to profile the microbiome through metagenomics and metatranscriptomics continue evolving, so does the definition. More recently this definition has evolved to include not only the microbes and their genomes but also the accumulation of environmental and host factors. Together the dynamic interaction between microbial genomes and host environments has significant implications on both health and disease risk. This conceptual shift is accompanied by the recognition that to achieve a more comprehensive microbiome study, metagenomics and metatranscriptomics must be combined with metabolomics. The metabolome is considered the most direct indicator of health, as it sits at the nexus of an organism's genetic blueprint and environmental stimuli. Metabolomics is the comprehensive analysis of all small molecules and has become a relevant technology for biomarker identification in a range of diseases. Moreover integrating metagenomics, metatranscriptomics, and metabolomics can lead

to deeper insights into biological pathways that govern host–microbial interaction.

Keywords

Metabolomics; sample preservation; dried blood spots; data processing; multiomics

Why metabolomics

The past decade has seen a revolutionary approach to metabolic research by advancements in analytical chemistry techniques such as mass spectrometry (MS) and nuclear magnetic resonance spectroscopy. These techniques facilitate the analysis of myriad chemical moieties that together form the metabolic complement of an organism. Development and adaptation of this approach have given rise to the field of metabolomics, the latest addition to the "omics" family (Fiehn, 2001; Riekeberg & Powers, 2017). Metabolomics is the identification of small molecules known as metabolites. Metabolomics is a powerful companion to other commonly used omics techniques: genomics, transcriptomics, and proteomics. While these techniques provide genetic and functional potential information, metabolomics goes a layer deeper by providing phenotypic information. Our genomes, transcriptomes, proteomes, and metabolomes do not exist in a vacuum—they impact and are impacted by one another as well as by the outside environment

Since its inception metabolomics has made remarkable progress in providing useful systematic insight into the mechanisms underlying numerous diseases, including type 2 diabetes (Qi et al., 2022), cardiovascular disease (Fromentin et al., 2022), and neurological disorders (Kong et al., 2021). This burgeoning area provides opportunities to paint a dynamic picture of an organism's phenotype through the comprehensive analysis of endogenous and exogenous metabolites in biological fluids, tissues, and cells.

While a number of definitions exist for metabolomics, it is favorable to use the definition of the Metabolomics Society, which interprets

Page 226 metabolomics as a “comprehensive and holistic identification and quantification of intracellular and extracellular metabolites in a biological system” (Goodacre et al., 2004; Nicholson et al., 2002). Metabolomics provides an overview of the biochemical processes in a given sample at a point in time. The metabolome is the complete set of small molecular mass metabolites (<1500 D) in diverse atomic arrangements in a biological system or the total complement of metabolites in the cell (Patti et al., 2012). These molecules can be either the substrates or products of the complex biochemical networks associated with the cellular and systemic biological pathways. Amino acids, fatty acids, carbohydrates, organic acids, lipids, and nucleotides are the common metabolites involved in various physiological processes for signaling, survival, energy production, inflammation, oxidative balance, and anabolic and catabolic reactions that result in cell and tissue growth (Dunn et al., 2005). Small molecules are biologically regulated through multiple processes including feedback inhibition by other small molecules, covalent modification (posttranslation modification) of proteins that activate or inhibit protein activity, and modification of transcripts (mRNA) and the proteins of the cellular pathways as a whole (Fiehn, 2002; Pearson, 2007). The ultimate goal of metabolomics is to quantify all of the metabolites in a physiological system in a given state and at a given point in time. However the number of metabolites and their enormous diversity make such quantification virtually impossible (Patti et al., 2012).

Targeted versus untargeted metabolomics

Metabolomics aims to establish various analytical methods for rapid, high-throughput, and comprehensive analysis. The most important part of any metabolomics study is defining the scope of your study. Two approaches may be leveraged: global profiling or targeted. Global profiling, otherwise known as untargeted metabolomics, is semiquantitative and considers all output metabolite data, while targeted studies utilize panels of known metabolites to

quantify the amounts of those specific metabolites in the sample(s) of interest (Schrimpe-Rutledge et al., 2016).

Targeted metabolomics is based on data hypothesis-driven research, which requires a predefined set of metabolites (Djukovic et al., 2013; Gao et al., 2018; Zhou & Yin, 2016). Typically the quantification and characterization of the metabolites are focused on selected metabolites and some related etiological pathways. The targeted analysis approach typically uses isotopically labeled internal standards, external calibration curves, and multiple reaction monitoring for absolute quantification in which the combination of chromatography and tandem MS is necessary (Roberts et al., 2012). Analytical peaks are usually identified and quantified based on a unique combination of chromatographic retention time (RT), mass precursor ion m/z ratio, and fragment ion m/z ratio (Dunn et al., 2011; Schrimpe-Rutledge et al., 2016). The data are scanned for specific compounds based on the identified m/z ratios and are subsequently characterized and quantified against a calibration standard of known concentration. The major advantage of this approach is that the absolute concentration of each targeted metabolite in a sample is determined.

A nontargeted or global metabolic profiling provides a hypothesis-free analysis of metabolites by incorporating the fields of chemistry and bioinformatics (Alonso et al., 2015). An untargeted approach provides an overview of all metabolites in a sample without a researcher having to narrow the focus on select metabolites. Another prominent advantage to global profiling is the potential identification of unknown metabolites not previously quantified and characterized. In spite of this the identification of new metabolites should be further validated in a targeted manner for confirmation (Chaleckis et al., 2019; Cho et al., 2014). To achieve these objectives the recent advances are focused on the development of new workflows for accurate metabolite annotation in untargeted analysis, data acquisition methods to realize high-throughput analysis, new derivatization and detection methods to cover more metabolites and libraries and

algorithms to facilitate data processing (Dunn et al., 2011; Goodacre et al., 2004). Hybrid approaches are starting to emerge, where synergetic approaches and statistical tools might provide dynamic phenotype pictures of the physiological system as a whole (Alonso et al., 2015; Cajka & Fiehn, 2016). These aim to combine the advantages of both untargeted and targeted metabolomics, gaining improved repeatability and a wider detection range (Cajka & Fiehn, 2016).

Sample handling considerations for metabolomics analyses

Sample handling methods play a crucial role in maintaining the quality of data acquired for metabolomics studies. The most important factor in terms of sample handling is the storage temperature after collection, as this can significantly impact the stability and integrity of the Page 228 metabolites present in the sample. Storage conditions can range from room temperature to −20°C to −80°C. Storing samples at room temperature is the least favorable option for long-term storage, as it can lead to high variability, interferences with instrumentation, and degradation of metabolites (Smith et al., 2020). This method is suitable for short-term storage or when immediate processing is not possible but should be validated to be suitable for the analytes of interest. Storing samples at −20°C provides a better preservation environment compared with room temperature. Surprisingly little data are available on the long-term effects of this temperature on sample stability, but it is generally considered a suitable option for short- to medium-term storage (Valo et al., 2022). Some metabolites may still degrade over time at this temperature. Storing samples at −80°C is the most effective method for long-term storage, as it minimizes degradation and maintains the integrity of the metabolites (Smith et al., 2020). This temperature is commonly used in metabolomics studies to ensure the highest data quality and the most complete metabolite profile (Chen et al., 2020). However it is also the most expensive and requires specialized equipment and infrastructure.

Another consideration is the type of sample being studied, which plays a significant role in determining appropriate collection and processing procedures. Whole blood samples can be analyzed intact to include the cellular constituents or may be processed into two primary components: plasma and serum. These two components are obtained by centrifugation separating the liquid portion of blood from its cellular constituents. Serum is the result of blood clotting, while plasma is obtained after the addition of an anticoagulant. Ethylenediaminetetraacetic acid (EDTA), citrate, and heparin are widely used anticoagulants for the collection of plasma (Stevens et al., 2019). Regardless of the blood type collection avoiding hemolysis and red blood cell destruction is paramount. Hemolysis releases hemoglobin and other intracellular components that can compromise the accurate quantification of other blood metabolites through MS. Furthermore they should be stored as aliquots in separate sample tubes to minimize the occurrence of freeze-thaw cycles, which have the potential to alter the composition of blood metabolites (Gao et al., 2018; Goodman et al., 2021).

Urine comprises water-soluble molecules, including urea, toxins, and various waste products extracted from the bloodstream by the kidneys. Furthermore its ease of collection, both at home and in laboratory settings, makes urine a valuable biological sample. Nonetheless certain

considerations are essential when handling urine samples. The first is that urine samples can contain elevated concentrations of salts, which may lead to adduct formation during MS analysis. The presence of multiple adducts can diminish sensitivity in metabolite detection. To address this sample extraction should effectively remove interfering salts, or their signals must be identified and subsequently removed during data analysis (Edmands et al., 2014). The second consideration is that highly abundant metabolites, such as urea, may suppress MS signals from low-abundance metabolites. In such cases urine may need to be diluted to enhance the signal strength of less prevalent metabolites. However this step requires a normalization process to account for variations in urine volume, water content, and water density

before and after dilution (Laparre et al., 2017). Lastly timely processing is crucial for the generation of robust urine metabolomes. Urine metabolomes can change rapidly, even when stored at temperatures as low as 4°C. If quick processing is unfeasible, storage at −20°C or −80°C can help maintain the integrity of urine metabolomes prior to processing (Emwas et al., 2015).

Fecal metabolomes serve as a valuable proxy for assessing the activities of the gut microbiome and their connections to host health, lifestyle, and genetic factors (Zierer et al., 2018). The intricate links between fecal metabolomes and various aspects of digestive and bodily health, including the gut-brain axis, make them a compelling subject of study (Konjevod et al., 2021). Analyzing fecal metabolomes presents a set of technical considerations that influence processing protocols. Fecal matter is a complex mixture of biomolecules. Despite being approximately 75% water fecal matter contains a diverse array of biomolecules, including undigested organic material, bacterial components, carbohydrates, lipids, proteins, and organic acids. Fecal matter exhibits high heterogeneity, and its appearance and chemical composition can be influenced by several factors, including diet, the health status of the individual, medication usage, and the individual's genetic background (Rose et al., 2015). This diversity complicates efforts to establish standardized extraction protocols for fecal matter (Lin et al., 2007).

Regarding sample storage fecal samples can be preserved at room temperature for a brief duration when using certain preservatives. Using the validated OMNI GUT tubes is one way to enable room-temperature storage of human fecal samples prior to processing, which preserves the metabolomic profile (Lim et al., 2020; Ramamoorthy et al., 2021). Page 230 However, for long-term storage, maintaining sample integrity necessitates storing them at −80°C.

Tissue

Exploring tissue metabolomes offers valuable insights for biomedical research. Tissue biopsies play a critical role in diagnosing chronic diseases,

studying organ physiology, and monitoring disease progression. Cellular metabolic processes continue even after tissue biopsies are collected (Saoi & Britz-McKibbin, 2021). Therefore it is essential to process samples promptly or, if not feasible, to flash-freeze them in liquid nitrogen. This preserves the tissue metabolome as closely as possible to its in situ state (Want et al., 2013). An alternative approach involves storing samples at −80°C until processing. Tissue samples may contain extraneous elements such as connective tissue, fat, and residual blood, which can obscure the tissue metabolome (Daskalaki et al., 2018). To prevent contamination any tissue extraction protocol must begin by removing these components through rinsing with deionized water. Just as with other sample types various organic solvents have been employed in the processing of tissue samples, under the assumption that metabolomes are uniform and reproducible. However different tissue types may necessitate unique extraction protocols (Rose et al., 2015). While progress has been made in standardizing these procedures, further refinement is required.

Dried blood spots

A dried blood spot (DBS) is a small volume of blood from a heel- or finger-prick collected onto filter paper and dried. DBS sampling is especially suitable for studying remote and underserved populations in resource-limited settings (Pfluger et al., 2022; Tobin et al., 2021), as they are cost-effective, minimally invasive, easy to collect, and can be transported and stored at room temperature with minimal negative impacts on a variety of biochemical classes (Goodman, 2023). In contrast with liquid samples (e.g., whole blood, plasma, and serum) DBS does not contain active enzymes that continually act on substrates in the sample at room temperature, as enzyme activity is halted once the sample is dried. However, similar to the negative effect of hemolysis in a plasma and serum sample, metabolite levels in a DBS sample are affected by the amount of red blood cells in the blood sample applied to the filter paper. Page 231

Sample extraction

Metabolite extraction is a critical step in metabolomics studies, as it directly impacts the quality and comprehensiveness of the metabolomic data obtained. Several extraction methods have been developed and optimized for various sample types and metabolite classes. Here is an overview of commonly used extraction methods in metabolomics:

Solvent-based extractions are widely employed due to their simplicity and effectiveness and use organic solvents, either alone or in combination, to disrupt cell membranes and extract metabolites. Common solvents include methanol, ethanol, acetonitrile, and chloroform. For example the methanol–chloroform–water extraction, also known as the Bligh and Dyer method, is frequently used for its ability to extract both polar and nonpolar metabolites. This method involves a biphasic separation, allowing for the simultaneous extraction of hydrophilic (polar) and lipophilic (nonpolar) compounds.

Solid-phase extraction (SPE) is another important technique, particularly useful for sample clean-up and metabolite enrichment. SPE utilizes solid adsorbents to selectively retain metabolites of interest while removing interfering compounds. Various SPE cartridges with different chemistries (e.g., reversed-phase, ion-exchange, or mixed-mode) are available, allowing for tailored extractions based on the physicochemical properties of the target metabolites. Additional steps may be required for tissue samples to disrupt the cellular structure before extraction. These can include mechanical homogenization, sonication, or freeze-thaw cycles. Adding liquid nitrogen to snap-freeze cells, followed by applying cold methanol, is an effective method to rapidly quench metabolic activity and extract intracellular metabolites. Extraction protocols may also be optimized for specific metabolite classes. For example lipid extractions often employ nonpolar solvents like chloroform or methyl tert-butyl ether, while polar metabolites are typically extracted using aqueous-organic solvent mixtures.

The choice of extraction method can significantly impact the metabolite coverage and the repeatability of the analysis. Factors such as sample type,

target metabolites, and the analytical platform to be used should be considered when selecting an extraction method. Furthermore method validation, including assessment of extraction efficiency, reproducibility, and potential matrix effects, is crucial for ensuring the reliability of metabolomic data. While there is no universal extraction method suitable for all metabolomic applications, Page 232 understanding the principles and limitations of different extraction techniques allows researchers to select and optimize protocols appropriate for their specific research questions and sample types.

What are the instruments used?

Different analytical platforms are employed to survey the vast chemical diversity and wide dynamic concentration ranges of metabolites (Kaczmarek et al., 2023; Nash & Dunn, 2019). MS is commonly coupled to separation methods such as liquid chromatography-mass spectrometry (LC-MS), gas chromatography-mass spectrometry (GC-MS), and capillary electrophoresis. While each chromatography method has strengths and weaknesses, LC-MS accounts for most metabolomics studies due to quicker and less extensive extraction procedures and the ability to cover a broader range of metabolites (Nash & Dunn, 2019). Liquid chromatography can be further divided into three methods based on polarity: reverse-phase liquid chromatography, hydrophilic interaction liquid chromatography (HILIC), and ion chromatography. For each type of chromatography the selection of column type, mobile phase solvents, flow rate, and gradient are all parameters that need to be optimized for robust chromatography and reliable LC-MS experiments (Kaczmarek et al., 2023).

Metabolite detection by mass analyzers with high resolution and sensitivity is generally desired. The resolution focuses on the mass accuracy of ions in the mass spectrum, allowing for the discrimination of closely related species. At the same time sensitivity pertains to the instrument's ability to detect and quantify ions, particularly at low concentrations. In full scan mode, also

known as MS^1, only one MS function is acquired to generate ions of the molecular species, adducts, and in-source fragments, giving rise to the mass-to-charge ratio (m/z) and relative abundance (Bauermeister et al., 2022; Defossez et al., 2023). MS^1 is used without fragmentation and is commonly used for high-throughput analysis due to its simplicity in acquisition, data processing, and high performance in discriminating biological samples. To improve data quality and metabolite annotation rate fragmentation of precursor ions (found in MS^1) is facilitated by MS/MS, which yields MS^2 data. Data acquisition is performed in positive and negative ionization modes to fully capture analytes that ionize more efficiently in either mode. Together the LC-MS dataset will report the following measurements: chromatographic RT, full scan m/z, MS/MS, and Page 233 relative abundance, culminating in the "high-dimensional dataset." These parameters combine to define the "molecular fingerprint" of a metabolite present in a sample.

Data processing and annotation

Metabolomics data processing is a critical step in extracting meaningful information from complex datasets. The process begins with the conversion of raw data from proprietary instrument formats into open, standardized formats such as mzML (Chang et al., 2021; Graça et al., 2022). This is followed by preprocessing steps, which for LC-MS data often include feature detection (peak-picking) in individual samples and subsequent feature matching across samples. These steps involve identifying and quantifying chromatographic peaks that represent distinct metabolites. Additional preprocessing may include RT alignment to correct for slight shifts between runs, noise filtering, and intensity normalization. The result is a data matrix with samples in one dimension and features (unique chromatographic peaks) in the other, ready for statistical analysis and biological interpretation (Misra, 2021; Zhou et al., 2012).

Metabolite annotation, the process of assigning putative identities to detected features, is a challenging step in metabolomics workflows.

Metabolite annotation remains the critical bottleneck in untargeted metabolomics data analysis and interpretation due to an “infinite chemical dark space” (Graça et al., 2022; Zulfiqar et al., 2023). The attribution of confidence levels to metabolites is contingent upon the chosen methodological framework delineated by the Metabolomics Standards Initiative (Schymanski et al., 2014; Sumner et al., 2007). Level 1 annotation represents the highest confidence, where at least two physiochemical properties of either m/z, RT, MS/MS, and collision cross-section must be matched with an authentic standard. The error rate for this type of identification is minimal, with stringent scoring methods and cutoff values for matching applied. A “Level 2” annotation relates to m/z from experimental data being searched against public/commercial reference spectral libraries, resulting in probable structure. To facilitate such comparisons extensive molecular structural MS (Riekeberg & Powers, 2017) databases (Kim et al., 2019), HMDB (Wishart et al., 2022), KEGG (Kanehisa et al., 2016), and ChemSpider (Pence & Williams, 2010), and MS^2 spectral databases, METLIN (Xue et al., 2020), GNPS (Nothias et al., 2020), MassBank (Horai et al., 2010), Page 234 and NIST (NIST Standard Reference Database 1A, 2014), are used. Emerging strategies based on in silico and network modeling based on the substructure of features are being utilized. These approaches yield a substantial number of annotations, and their identification tends to align with Level 3 and 4 annotations. Level 3 annotation refers to putatively characterized compound classes, where the general chemical family is tentatively identified based on partial structural information. Level 4 annotation, in contrast, denotes unknown compounds that can be detected and measured but lack any structural characterization or class identification. Well-known in silico tools are CFM-ID (Wang et al., 2021), MetFrag (Ruttkies et al., 2016), MS-FINDER (Tsugawa et al., 2016), SIRIUS (Dührkop et al., 2019), CANOPUS, and COSMIC (Hoffmann et al., 2022). While network-based and in silico approaches facilitate a quicker route to metabolite annotation, they have a higher error rate compared to Level 1

annotations. Therefore downstream biological interpretation and its integration with other data streams should caveat the use of Level 2–4 metabolite annotation.

Integration of metabolomics and metagenomics

The previous sections have provided an overview of the generation of metabolomics data. This differs significantly from generating metagenomics and metatranscriptomics data, which are produced by sequencing technologies. One massive challenge is to integrate these heterogeneous datasets to elucidate the large, dynamic, and complex network of interactions between each constituent entity (Aguiar-Pulido et al., 2016). Advanced computational techniques, such as machine learning and network analysis, are increasingly employed to uncover hidden patterns and interactions among the different data types (**Fig. 11.1**) (Reverter et al., 2020). For instance correlating metagenomic data with metabolomic profiles can reveal how specific microbial taxa influence metabolite production, while linking metatranscriptomic data can help identify which genes are responsible for these metabolic changes. However, while this interplay of datasets can infer such insights, it remains difficult to ascertain whether a metabolite was generated by the host or by the microbiome (Aguiar-Pulido et al., 2016). In this context creating host metabolomes and microbially derived signatures can be characterized from matrices such as fecal and plasma/serum (Psychogios, 2011; Zierer et al., 2018). Using the fecal metabolome and metagenome can inform the metabolic interplay

Page 235 Page 236

between the host, diet, and gut microbiota, while the host metabolome can reveal how commensal microbiota impacts other biological systems (Puig-Castellví et al., 2023).

Sampling
Research Question?

Chemical Extraction

DNA Extraction

LC-MS/MS

16S rRNA Sequencing

Long-Read Sequencing

Chemometrics Feature Selection

Numerical Analysis Operational taxonomic Unit (OTU) Selection

Correlation Network

Metabolites correlated with OTUs (corr > 0.8)

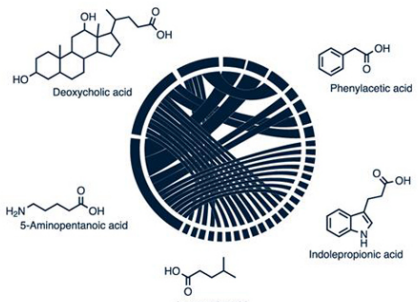


FIGURE 11.1 An overview of the integration of metabolomics and metagenomics tools.

Follow for extended description

Conclusion and future directions

In this chapter we have discussed metabolomics and the need to incorporate it when engaging in microbiome analysis to provide a more comprehensive understanding. However the main takeaway of this chapter is the difference in data generation between metagenomics and metabolomics. It is important to note that sample collection and storage, sample extraction, and the technologies used (sequencing vs MS) are all drastically different. However all datasets produced are inherently complex and noisy, which requires optimal protocols that lead to clean datasets that can be integrated. It is important that these are defined early to minimize user variation and allow for comparison of biological differences. Altogether the integrative approach of the host environment with the microbiome has the potential to elucidate biological mechanisms that govern health and disease risk, specifically in diseases where the different microbiomes of the human body have a close interaction with host tissues (e.g., Crohn's disease/inflammatory bowel diseases and colon cancer).

References

Aguiar-Pulido et al., 2016 Aguiar-Pulido V, Huang W, Suarez-Ulloa V, Cickovski T, Mathee K, Narasimhan G. Metagenomics, metatranscriptomics, and metabolomics approaches for microbiome analysis. *Evolutionary Bioinformatics*. 2016;12:5–16 **https://doi.org/10.4137/EBO.S36436**.

Alonso et al., 2015 Alonso A, Marsal S, Julià A. Analytical methods in untargeted metabolomics: State of the art in 2015. *Frontiers in*

Bioengineering and Biotechnology. 2015;3 **https://doi.org/10.3389/fbioe.2015.00023**.

Bauermeister et al., 2022 Bauermeister A, Mannochio-Russo H, Costa-Lotufo LV, Jarmusch AK, Dorrestein PC. Mass spectrometry-based metabolomics in microbiome investigations. *Nature Reviews Microbiology*. 2022;20(3):143–160 **https://doi.org/10.1038/s41579-021-00621-9**.

Cajka and Fiehn, 2016 Cajka T, Fiehn O. Toward merging untargeted and targeted methods in mass spectrometry-based metabolomics and lipidomics. *Analytical Chemistry*. 2016;88(1):524–545 **https://doi.org/10.1021/acs.analchem.5b04491**.

Chaleckis et al., 2019 Chaleckis R, Meister I, Zhang P, Wheelock CE. Challenges, progress and promises of metabolite annotation for LC–MS-based metabolomics. *Current Opinion in Biotechnology*. 2019;55:44–50 **https://doi.org/10.1016/j.copbio.2018.07.010**.

Chang et al., 2021 Chang HY, Colby SM, Du X, et al. A practical guide to metabolomics software development. *Analytical Chemistry*. 2021;93(4):1912–1923 **https://doi.org/10.1021/acs.analchem.0c03581**.

Chen et al., 2020 Chen D, Han W, Huan T, Li L, Li L. Effects of freeze–thaw cycles of blood samples on high-coverage quantitative metabolomics. *Analytical Chemistry*. 2020;92(13):9265–9272 **https://doi.org/10.1021/acs.analchem.0c01610**.

Cho et al., 2014 Cho K, Mahieu NG, Johnson SL, Patti GJ. After the feature presentation: Technologies bridging untargeted metabolomics and biology. *Current Opinion in Biotechnology*. 2014;28:143–148 **https://doi.org/10.1016/j.copbio.2014.04.006**.

Daskalaki et al., 2018 Daskalaki E, Pillon NJ, Krook A, Wheelock CE, Checa A. The influence of culture media upon observed cell secretome metabolite profiles: The balance between cell viability and data interpretability. *Analytica Chimica Acta*. 2018;1037:338–350 **https://doi.org/10.1016/j.aca.2018.04.034**.

Defossez et al., 2023 Defossez E, Bourquin J, von Reuss S, Rasmann S, Glauser G. Eight key rules for successful data-dependent acquisition in mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*. 2023;42(1):131–143 **https://doi.org/10.1002/mas.21715**.

Djukovic et al., 2013 Djukovic D, Gowda GAN, Raftery D. *Mass spectrometry and NMR spectroscopy-based quantitative metabolomics*. *Proteomic and Metabolomic Approaches to Biomarker Discovery* United States: Elsevier Inc.; 2013;279–297 **https://doi.org/10.1016/B978-0-12-394446-7.00018-2**.

Dührkop et al., 2019 Dührkop K, Fleischauer M, Ludwig M, et al. SIRIUS 4: A rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods*. 2019;16(4):299–302 **https://doi.org/10.1038/s41592-019-0344-8**.

Dunn et al., 2005 Dunn WB, Bailey NJC, Johnson HE. Measuring the metabolome: Current analytical technologies. *Analyst*. 2005;130(5):606–625 **https://doi.org/10.1039/b418288j**.

Dunn et al., 2011 Dunn WB, Broadhurst D, Begley P, et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols*. 2011;6(7):1060–1083 **https://doi.org/10.1038/nprot.2011.335**.

Edmands et al., 2014 Edmands WMB, Ferrari P, Scalbert A. Normalization to specific gravity prior to analysis improves

information recovery from high resolution mass spectrometry metabolomic profiles of human urine. *Analytical Chemistry*. 2014;86(21):10925–10931 **https://doi.org/10.1021/ac503190m**.

Emwas et al., 2015 Emwas AH, Luchinat C, Turano P, et al. Standardizing the experimental conditions for using urine in NMR-based metabolomic studies with a particular focus on diagnostic studies: A review. *Metabolomics: Official Journal of the Metabolomic Society*. 2015;11(4):872–894 **https://doi.org/10.1007/s11306-014-0746-7**.

Fiehn, 2002 Fiehn O. Metabolomics—The link between genotypes and phenotypes. *Plant Molecular Biology*. 2002;48(1-2):155–171 **https://doi.org/10.1023/A:1013713905833**.

Fiehn, 2001 Fiehn O. Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comparative and Functional Genomics*. 2001;2(3):155–168 **https://doi.org/10.1002/cfg.82**.

Fromentin et al., 2022 Fromentin S, Forslund SK, Chechi K, et al.... Microbiome and metabolome features of the cardiometabolic disease spectrum. *Nature Medicine*. 2022;28(2):303–314 **https://doi.org/10.1038/s41591-022-01688-4**.

Gao et al., 2018 Gao Y, Chen Y, Yue X, et al. Development of simultaneous targeted metabolite quantification and untargeted metabolomics strategy using dual-column liquid chromatography coupled with tandem mass spectrometry. *Analytica Chimica Acta*. 2018;1037:369–379 **https://doi.org/10.1016/j.aca.2018.08.042**.

Goodacre et al., 2004 Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB. Metabolomics by numbers: Acquiring and understanding global metabolite data. *Trends in Biotechnology*.

2004;22(5):245–252 **https://doi.org/10.1016/j.tibtech.2004.03.007**.

Goodman, 2023 Goodman KD. Dried blood spots as a suitable matrix for phenotyping the metabolome: A proof of concept study. *Precision Medicine Q*. 2023;1.

Goodman et al., 2021 Goodman K, Mitchell M, Evans AM, et al. Assessment of the effects of repeated freeze thawing and extended bench top processing of plasma samples using untargeted metabolomics. *Metabolomics: Official Journal of the Metabolomic Society*. 2021;17 **https://doi.org/10.1007/s11306-021-01782-7**.

Graça et al., 2022 Graça G, Cai Y, Lau C-HE, et al. Automated annotation of untargeted all-ion fragmentation LC–MS metabolomics data with MetaboAnnotatoR. *Analytical Chemistry*. 2022;94(8):3446–3455 **https://doi.org/10.1021/acs.analchem.1c03032**.

Hoffmann et al., 2022 Hoffmann MA, Nothias LF, Ludwig M, et al. High-confidence structural annotation of metabolites absent from spectral libraries. *Nature Biotechnology*. 2022;40(3):411–421 **https://doi.org/10.1038/s41587-021-01045-9**.

Horai et al., 2010 Horai H, Arita M, Kanaya S, et al.... MassBank: A public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*. 2010;45(7):703–714 **https://doi.org/10.1002/jms.1777**.

Kaczmarek et al., 2023 Kaczmarek M, Zhang N, Buzhansky L, Gilead S, Gazit E. Optimization strategies for mass spectrometry-based untargeted metabolomics analysis of small polar molecules in human plasma. *Metabolites*. 2023;13(8):923 **https://doi.org/10.3390/metabo13080923**.

Kanehisa et al., 2016 Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein

annotation. *Nucleic Acids Research*. 2016;44(D1):D457 **https://doi.org/10.1093/nar/gkv1070**.

Kim et al., 2019 Kim S, Chen J, Cheng T, et al. PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Research*. 2019;47(1):D1102 **https://doi.org/10.1093/nar/gky1033**.

Kong et al., 2021 Kong G, Ellul S, Narayana VK, et al. An integrated metagenomics and metabolomics approach implicates the microbiota-gut-brain axis in the pathogenesis of Huntington's disease. *Neurobiology of Disease*. 2021;148 **https://doi.org/10.1016/j.nbd.2020.105199**.

Konjevod et al., 2021 Konjevod M, Perkovic MN, Sáiz J, Strac DS, Barbas C, Rojo D. Metabolomics analysis of microbiota-gut-brain axis in neurodegenerative and psychiatric diseases. *Journal of Pharmaceutical and Biomedical Analysis*. 2021;194 **https://doi.org/10.1016/j.jpba.2020.113681**.

Laparre et al., 2017 Laparre J, Kaabia Z, Mooney M, et al. Impact of storage conditions on the urinary metabolomics fingerprint. *Analytica Chimica Acta*. 2017;951:99–107 **https://doi.org/10.1016/j.aca.2016.11.055**.

Lim et al., 2020 Lim MY, Hong S, Kim B-M, Ahn Y, Kim H-J, Nam Y-D. Changes in microbiome and metabolomic profiles of fecal samples stored with stabilizing solution at room temperature: A pilot study. *Scientific Reports*. 2020;10(1):1789 **https://doi.org/10.1038/s41598-020-58719-8**.

Lin et al., 2007 Lin CY, Wu H, Tjeerdema RS, Viant MR. Evaluation of metabolite extraction strategies from tissue samples using NMR metabolomics. *Metabolomics: Official Journal of the Metabolomic Society*. 2007;3(1):55–67 **https://doi.org/10.1007/s11306-006-0043-1**.

Misra, 2021 Misra BB. New software tools, databases, and resources in metabolomics: Updates from 2020. *Metabolomics: Official*

Journal of the Metabolomic Society. 2021;17 **https://doi.org/10.1007/s11306-021-01796-1**.

Nash and Dunn, 2019 Nash WJ, Dunn WB. From mass to metabolite in human untargeted metabolomics: Recent advances in annotation of metabolites applying liquid chromatography-mass spectrometry data. *TrAC—Trends in Analytical Chemistry*. 2019;120 **https://doi.org/10.1016/j.trac.2018.11.022**.

Nicholson et al., 2002 Nicholson JK, Connelly J, Lindon JC, Holmes E. Metabonomics: A platform for studying drug toxicity and gene function. *Nature Reviews Drug Discovery*. 2002;1(2):153–161 **https://doi.org/10.1038/nrd728**.

NIST Standard Reference Database 1A, 2014 NIST Standard Reference Database 1A. NIST. (2014).

Nothias et al., 2020 Nothias LF, Petras D, Schmid R, et al.... Feature-based molecular networking in the GNPS analysis environment. *Nature Methods*. 2020;17(9):905–908 **https://doi.org/10.1038/s41592-020-0933-6**.

Patti et al., 2012 Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: The apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology*. 2012;13(4):263–269 **https://doi.org/10.1038/nrm3314**.

Pearson, 2007 Pearson H. Meet the human metabolome. *Nature*. 2007;446(7131):8 **https://doi.org/10.1038/446008a**.

Pence and Williams, 2010 Pence HE, Williams A. Chemspider: An online chemical information resource. *Journal of Chemical Education*. 2010;87(11):1123–1124 **https://doi.org/10.1021/ed100697w**.

Pfluger et al., 2022 Pfluger BA, Smith HV, Weber AM, et al. Non-targeted dried blood spot-based metabolomics analysis showed rice bran supplementation effects multiple metabolic pathways

during infant weaning and growth in Mali. *Nutrients*. 2022;14(3):609 **https://doi.org/10.3390/nu14030609**.

Psychogios, 2011 Psychogios N. The human serum metabolome. *PLoS One*. 2011;6.

Puig-Castellví et al., 2023 Puig-Castellví F, Pacheco-Tapia R, Deslande M, et al. Advances in the integration of metabolomics and metagenomics for human gut microbiome and their clinical applications. *TrAC Trends in Analytical Chemistry*. 2023;167 **https://doi.org/10.1016/j.trac.2023.117248**.

Qi et al., 2022 Qi Q, Li J, Yu B, et al.... Host and gut microbial tryptophan metabolism and type 2 diabetes: An integrative analysis of host genetics, diet, gut microbiome and circulating metabolites in cohort studies. *Gut*. 2022;71(6):1095–1105 **https://doi.org/10.1136/gutjnl-2021-324053**.

Ramamoorthy et al., 2021 Ramamoorthy S, Levy S, Mohamed M, et al. An ambient-temperature storage and stabilization device performs comparably to flash-frozen collection for stool metabolomics in infants. *BMC Microbiology*. 2021;21(1):59 **https://doi.org/10.1186/s12866-021-02104-6**.

Reverter et al., 2020 Reverter M, Rohde S, Parchemin C, Tapissier-Bontemps N, Schupp PJ. Marine metabolomics, biotechnology, coupling metabolite profiling and organism biology for the discovery of new compounds. *Front Mar Sci*. 2020;7613471 **https://doi.org/10.3389/fmars.2020.613471**.

Riekeberg and Powers, 2017 Riekeberg E, Powers R. New frontiers in metabolomics: From measurement to insight. *F1000Research*. 2017;6:1148 **https://doi.org/10.12688/f1000research.11495.1**.

Roberts et al., 2012 Roberts LD, Souza AL, Gerszten RE, Clish CB. Targeted metabolomics. *Current Protocols in Molecular Biology*. 2012;1 **https://doi.org/10.1002/0471142727.mb3002s98**.

Rose et al., 2015 Rose C, Parker A, Jefferson B, Cartmell E. The characterization of feces and urine: A review of the literature to inform advanced treatment technology. *Critical Reviews in Environmental Science and Technology*. 2015;45(17):1827–1879 **https://doi.org/10.1080/10643389.2014.1000761**.

Ruttkies et al., 2016 Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S. MetFrag relaunched: Incorporating strategies beyond in silico fragmentation. *Journal of Cheminformatics*. 2016;8 **https://doi.org/10.1186/s13321-016-0115-9**.

Saoi and Britz-McKibbin, 2021 Saoi M, Britz-McKibbin P. New advances in tissue metabolomics: A review. *Metabolites*. 2021;11672 **https://doi.org/10.3390/metabo11100672**.

Schrimpe-Rutledge et al., 2016 Schrimpe-Rutledge AC, Codreanu SG, Sherrod SD, McLean JA. Untargeted metabolomics strategies—challenges and emerging directions. *Journal of the American Society for Mass Spectrometry*. 2016;27(12):1897–1905 **https://doi.org/10.1007/s13361-016-1469-y**.

Schymanski et al., 2014 Schymanski EL, Jeon J, Gulde R, et al. Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environmental Science and Technology*. 2014;48(4):2097–2098 **https://doi.org/10.1021/es5002105**.

Smith et al., 2020 Smith L, Villaret-Cazadamont J, Claus SP, et al. Important considerations for sample collection in metabolomics studies with a special focus on applications to liver functions. *Metabolites*. 2020;10(3):104 **https://doi.org/10.3390/metabo10030104**.

Stevens et al., 2019 Stevens VL, Hoover E, Wang Y, Zanetti KA. Pre-analytical factors that affect metabolite stability in human urine, plasma, and serum: A review. *Metabolites*. 2019;9 **https://doi.org/10.3390/metabo9080156**.

Sumner et al., 2007 Sumner LW, Amberg A, Barrett D, et al.... Proposed minimum reporting standards for chemical analysis. *Metabolomics: Official Journal of the Metabolomic Society*. 2007;3(3):211–221 **https://doi.org/10.1007/s11306-007-0082-2**.

Tobin et al., 2021 Tobin NH, Murphy A, Li F, et al. Comparison of dried blood spot and plasma sampling for untargeted metabolomics. *Metabolomics: Official Journal of the Metabolomic Society*. 2021;17 **https://doi.org/10.1007/s11306-021-01813-3**.

Tsugawa et al., 2016 Tsugawa H, Kind T, Nakabayashi R, et al. Hydrogen rearrangement rules: Computational MS/MS fragmentation and structure elucidation using MS-FINDER software. *Analytical Chemistry*. 2016;88(16):7946–7958 **https://doi.org/10.1021/acs.analchem.6b00770**.

Valo et al., 2022 Valo E, Colombo M, Sandholm N, et al. Effect of serum sample storage temperature on metabolomic and proteomic biomarkers. *Scientific Reports*. 2022;12 **https://doi.org/10.1038/s41598-022-08429-0**.

Wang et al., 2021 Wang F, Liigand J, Tian S, Arndt D, Greiner R, Wishart DS. CFM-ID 4.0: More accurate ESI-MS/MS spectral prediction and compound identification. *Analytical Chemistry*. 2021;93(34):11692–11700 **https://doi.org/10.1021/acs.analchem.1c01465**.

Want et al., 2013 Want EJ, Masson P, Michopoulos F, et al. Global metabolic profiling of animal and human tissues via UPLC-MS. *Nature Protocols*. 2013;8(1):17–32 **https://doi.org/10.1038/nprot.2012.135**.

Wishart et al., 2022 Wishart DS, Guo AC, Oler E, et al.... HMDB 5.0: The Human Metabolome Database for 2022. *Nucleic Acids*

Research. 2022;50(1):D622 **https://doi.org/10.1093/nar/gkab1062**.

Xue et al., 2020 Xue J, Guijas C, Benton HP, Warth B, Siuzdak G. METLIN MS2 molecular standards database: A broad chemical and biological resource. *Nature Methods*. 2020;17(10):953–954 **https://doi.org/10.1038/s41592-020-0942-5**.

Zhou et al., 2012 Zhou B, Xiao JF, Tuli L, Ressom HW. LC-MS-based metabolomics. *Molecular Biosystems*. 2012;8(2):470–481 **https://doi.org/10.1039/C1MB05350G**.

Zhou and Yin, 2016 Zhou J, Yin Y. Strategies for large-scale targeted metabolomics quantification by liquid chromatography-mass spectrometry. *The Analyst*. 2016;141(23):6362–6373 **https://doi.org/10.1039/C6AN01753C**.

Zierer et al., 2018 Zierer J, Jackson MA, Kastenmüller G, et al. The fecal metabolome as a functional readout of the gut microbiome. *Nature Genetics*. 2018;50(6):790–795 **https://doi.org/10.1038/s41588-018-0135-7**.

Zulfiqar et al., 2023 Zulfiqar M, Gadelha L, Steinbeck C, Sorokina M, Peters K. MAW: The reproducible Metabolome Annotation Workflow for untargeted tandem mass spectrometry. *Journal of Cheminformatics*. 2023;15 **https://doi.org/10.1186/s13321-023-00695-y**.

Description of Figure 11.1

Sampling
Research Question?

Chemical Extraction

DNA Extraction

LC-MS/MS

16S rRNA Sequencing

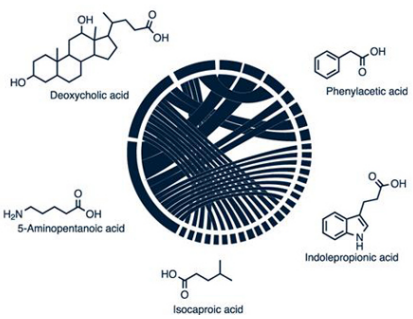
Long-Read Sequencing

Chemometrics Feature Selection

Numerical Analysis Operational taxonomic Unit (OTU) Selection

Correlation Network

Metabolites correlated with OTUs (corr > 0.8)



"Samples undergo extraction for metabolomics and DNA extraction for 16S rRNA and long-read sequencing. Each method has an independent data processing workflow for cleaning and normalization. Once high-quality datasets are obtained, statistical analyses, such as correlation" analysis ($r>0.8$), are applied to identify associations between metabolites and microbial taxa

Navigate back to Figure 11.1

CHAPTER 12

The event of multiomics in metagenomics

Himel Mallick[1,2,3], Mainak Bardhan[4,5], Chaitali Dagli[6], Ziyu Liu[1] and Jialin Gao[1], [1]Division of Biostatistics, Department of Population Health Sciences, Weill Cornell Medicine, Cornell University, New York, NY, United States, [2]Department of Statistics and Data Science, Cornell University, Ithaca, NY, United States, [3]Division of Gastroenterology and Hepatology, Weill Cornell Medicine, Cornell University, New York, NY, United States, [4]The John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, FL, United States, [5]The Dr. John T. Macdonald Foundation, Department of Human Genetics, University of Miami Miller School of Medicine, Miami, FL, United States, [6]Department of Epidemiology, School of Public Health, University of Alabama at Birmingham, Birmingham, AL, United States

Abstract

Microbiome research has made significant strides in recent years, transitioning from observational surveys to population-scale studies and mechanistic investigations. While substantial progress has been made in identifying individual microbes associated with outcomes, the molecular mechanisms by which microbes exert these effects remain poorly understood. Systems biology approaches leveraging multiomics have emerged to address this gap, aiming to provide a comprehensive understanding of host–microbiome molecular interactions by integrating multiple omics modalities. Microbial multiomics is inherently complex, requiring specialized statistical methods and computational tools for efficient analysis. In this chapter we first describe the primary omics technologies relevant to microbiome research. Next we survey techniques for analyzing and integrating microbiome multiomics data, including supervised machine learning, unsupervised approaches, and multimodal artificial intelligence. We also provide a set of case studies illustrating key scientific questions that investigators are addressing with microbiome multiomics. Finally we conclude with a discussion of future directions in this rapidly evolving field.

Keywords

Multiomics; metagenomics; metabolomics; metatranscriptomics; metaproteomics; systems biology; multimodal integration; multiview; human microbiome; multimodal AI

Introduction

The microbiome is a complex and dynamic ecosystem shaped by intricate interactions among microbes, the host, and the environment (Ma et al., 2024). These interactions evolve continuously under the influence of developmental stages, diet, lifestyle, and medical interventions (Gilbert et al., 2018). While metagenomics has significantly advanced our ability to catalog microbial diversity and functional potential, it cannot fully capture the multifactorial processes driving these interactions. To address this gap multiomics approaches have emerged, integrating data from metagenomics, metatranscriptomics, metabolomics, and metaproteomics to provide a comprehensive view of microbial communities (Chetty and Blekhman, 2024; Deek et al., 2024; Mallick et al., 2017). This integration allows researchers to unravel complex multimodal relationships, identify functional biomarkers, and link microbial activities to outcomes.

For instance metagenomics identifies the microbial species present in a sample but does not reveal which genes are actively expressed or the biochemical outputs influencing host biology. Metatranscriptomics complements this by highlighting gene expression patterns, while metabolomics uncovers microbial metabolites such as short-chain fatty acids (SCFAs) and bile acids (BAs), which regulate host metabolism and immune responses (Fleishman & Kumar, 2024; Morrison & Preston, 2016). Combining these layers of information generates Page 244 insights unattainable

through single-omics approaches alone (Beale et al., 2016). A key example is the Integrative Human Microbiome Project [iHMP (iHMP Research Network Consortium, 2014; Proctor et al., 2019)], which used multiomics to characterize microbiome correlates in pregnancy, inflammatory bowel diseases (IBD), and prediabetes. By uncovering patterns in microbial taxonomic, functional, and metabolite profiles of large human populations, the iHMP revealed dynamic functional changes within microbial communities, providing deeper insights into disease mechanisms and advancing our understanding of host–microbe interactions (Fettweis et al., 2019; Lloyd-Price et al., 2019; Zhou et al., 2019).

Beyond human health microbiome multiomics has further shaped ecological research (Beale et al., 2016). For example a recent multiomics study from the Earth Microbiome Project integrated amplicon sequencing, shotgun metagenomics, and untargeted metabolomics to examine 880 samples from terrestrial, marine, and host-associated habitats (Shaffer et al., 2022). The study revealed habitat-specific patterns linking microbial taxa and metabolites, with key species like *Conexibacter woesei* (terrestrial soils) and *Haloquadratum walsbyi* (marine deposits) associated with distinct metabolite profiles. Similarly a 2024 study applied multiomics and metaproteomics to marine plastic debris, identifying active hydrocarbonoclastic genera like *Psychrobacter and Pseudomonas*, along with expressed proteins involved in stress response, xenobiotic degradation, and plastic biodegradation (Messer et al., 2024).

Multiomics approaches thus provide complementary data streams useful for comprehensive biological investigation. The translational application of multiomics at the population scale, however, requires careful experimental, computational, and statistical considerations (Mallick et al., 2017; Miao et al., 2021; Tarazona et al., 2021). First identifying relevant biological mechanisms in multiomics studies demands high-resolution technologies and well-powered experimental designs—both challenging under current protocols. Second ensuring that multiomics signals are sufficiently reproducible for clinical actionability requires the development and adoption of multimodal tools, which are still in their infancy. Third standard single-omics tools from the literature cannot be directly applied to multiomics data without risking false positive or false negative results. This is particularly true because each omics modality has distinct data characteristics along with platform-specific preprocessing and quality control steps that Page 245 must be carefully considered to mitigate unique biases introduced during extraction, processing, and profiling. These challenges are further exacerbated by the technical nature of the associated data, which are typically noisy, heterogeneous, high-dimensional, and compositional, often with low signal-to-noise ratios, requiring augmentation of off-the-shelf tools (Ma et al., 2024; Mallick et al., 2021, 2024; Nickols, 2024).

In this chapter we outline current best practices and review available methods and computational tools for multiomics integration. We begin with an overview of key omics techniques used in multiomics studies, followed by case studies highlighting their diverse applications. We then explore computational frameworks for integrated multiomics analysis, including machine learning methods and multimodal artificial intelligence (AI). We conclude with a forward-looking perspective on the future directions for advancing microbiome research through integrated multiomics studies.

Multiomics approaches in metagenomics

In recent years microbiome research has moved beyond metagenomics, shifting focus from identifying "*who is there*" to understanding "*what they are doing*" and "*how they impact a biological outcome of interest*." Techniques like metatranscriptomics, metabolomics, and metaproteomics now provide deeper insights into microbial activity, metabolic outputs, and protein functions, respectively. In this section we briefly discuss these approaches. Because of space limitations we cannot fully explore these topics here but want the reader to understand their importance in uncovering the functional architecture of microbial communities and their broader biological significance.

Metabolomics

Metagenomics provides comprehensive insights into the genomic content of microbial communities, but it often falls short of revealing the functional outcomes of microbial activity. This is where metabolomics becomes essential,

bridging the gap between microbes and the biochemical products they produce (Beale et al., 2016). Metabolomics focuses on identifying and quantifying metabolites—small molecules that are produced as intermediates or end products of microbial metabolism. Metabolites such as SCFAs, BAs, amino acids, and lipids represent the functional output of microbial metabolism. For instance Page 246 SCFAs, including acetate, propionate, and butyrate, are produced by gut bacteria through the fermentation of dietary fibers. These molecules have significant effects on host physiology, influencing energy homeostasis, immune regulation, and intestinal integrity (Mann et al., 2024; Morrison & Preston, 2016; Ratajczak et al., 2019). BAs, which are modified by gut microbes through deconjugation and secondary transformation, serve as signaling molecules that regulate lipid metabolism, glucose homeostasis, and inflammation in the host (Collins et al., 2023; Fleishman & Kumar, 2024; Larabi et al., 2023). Sphingolipids, on the contrary, are bioactive lipids involved in inflammatory processes and have been linked to insulin resistance; they act as key players in the development of metabolic syndromes (Heaver et al., 2018). Similarly microbial metabolism of amino acids, such as tryptophan, produces derivatives that modulate neurotransmission, immune responses, and host–microbe signaling pathways (Ala, 2022; Roager & Licht, 2018). These metabolites not only provide insight into disease mechanisms but also offer therapeutic potential, especially in the context of microbe-associated metabolic disorders (Krautkramer et al., 2021; Saha et al., 2016).

Untargeted liquid chromatography - mass spectrometry (LC-MS) has emerged as a key tool in microbial metabolomics (Franzosa et al., 2019; Lloyd-Price et al., 2019) (**Table 12.1**). Its sensitivity and versatility enable the detection of a wide range of metabolites, including polar, nonpolar, and low-abundance compounds (Bauermeister et al., 2022). Untargeted approach is particularly valuable because it does not require prior knowledge of the sample composition and allows for the identification of novel metabolites, which are especially relevant in diverse microbial communities (Franzosa et al., 2019; Lloyd-Price et al., 2019). The global metabolic snapshot of microbial activity provided by LC-MS not only underscores the functional significance of microbial metabolites in disease progression but also opens doors to discovering previously uncharted biochemical pathways (Bauermeister et al., 2022; Beale et al., 2016). Beyond untargeted LC-MS other analytical techniques, such as gas chromatography - mass spectrometry (GC-MS) and nuclear magnetic resonance (NMR) spectroscopy, play pivotal roles in complementing metabolomic analyses (Bauermeister et al., 2022; Chen et al., 2019; Gallagher et al., 2021; Karu et al., 2018; Luan et al., 2019; Misheva et al., 2021; Smirnov et al., 2016; Zeki et al., 2020; Zhou et al., 2021). GC-MS is particularly effective in profiling volatile and thermally stable metabolites, expanding the metabolic Page 247 Page 248 scope beyond what LC-MS can achieve (Zeki et al., 2020). NMR spectroscopy, with its ability to deliver precise structural information, is invaluable for validating findings and integrating results with other omics modalities (Danzi et al., 2023). Together these methods strengthen the interpretability and depth of metabolomic studies, facilitating a more comprehensive understanding of microbial metabolic functions.

Table 12.1

Feature	Untargeted LC-MS	Targeted LC-MS	GC-MS	NMR spectroscopy
Scope	Wide range of metabolites	Specific metabolites	Volatile and thermally stable metabolites	Comprehensive structural information
Precision and quantification	Lower	Higher	Higher	Higher
Sample preparation	Minimal	Minimal	Requires chemical derivatization	Minimal
Data analysis	Complex	Relatively simpler	Relatively simpler	Relatively simpler
Sensitivity	Higher	Lower	Higher	Lower
Metabolite detection	Broad	Limited	Broad (with derivatization)	Narrower
Structural information	Limited	Limited	Limited	Detailed
Challenges	Data complexity, metabolite identification	Limited scope	Sample preparation, metabolite degradation	Sensitivity, instrument costs
Additional notes	Ideal for discovering unknown compounds	Beneficial for hypothesis-driven studies	Used in tandem with LC-MS for broader coverage	Provides precise structural profiles without standards

Metatranscriptomics

While metabolomics reveals the biochemical outputs of microbial activity, metatranscriptomics shifts the focus to gene expression, providing insights into the active functional processes occurring within microbial communities (Abu-Ali et al., 2018; Mehta et al., 2018). By analyzing total RNA metatranscriptomics enables researchers to identify which genes are actively transcribed and under what conditions these functional responses occur. The metatranscriptomic workflow involves RNA extraction from microbial communities, rRNA depletion or mRNA enrichment, cDNA synthesis and sequencing, and bioinformatic analysis of the resulting data (Lavelle & Sokol, 2018; Zhang, Thompson, Branck, et al., 2021; Zhang, Thompson, Huttenhower, et al., 2021). This technique has been successfully applied to diverse microbiomes (Abu-Ali et al., 2018; Lavelle & Sokol, 2018; Mehta et al., 2018; Schirmer et al., 2018; Zhang, Thompson, Branck, et al., 2021; Zhang, Thompson, Huttenhower, et al., 2021). Going beyond taxonomic composition metatranscriptomics reveals active but potentially low-abundance microbes, uncovers microbial responses to environmental stimuli, and highlights dynamic processes occurring in health and disease (Lavelle & Sokol, 2018). However the technique also poses challenges, including the low abundance of microbial mRNA relative to host RNA, the short half-life of bacterial mRNA, and the need for specialized protocols to deplete rRNA effectively (Zhang, Thompson, Huttenhower, et al., 2021). With ongoing technological improvements its utility in microbiome research continues to expand, enhancing our understanding of microbial implications in health and disease.

Metaproteomics

In contrast to metatranscriptomics, which captures microbial gene expression, metaproteomics provides a direct measure of the functional proteins produced, offering a closer view of the biochemical machinery Page 249 driving microbial activity (Sun et al., 2024). Metaproteomics begins with protein extraction from microbiome samples, which can include body fluids like feces, saliva, or tissue samples (Hickl et al., 2019; Karaduta et al., 2021). Extracted proteins are enzymatically digested into smaller fragments known as peptides, which are then analyzed using mass spectrometry (MS) to identify and quantify the proteins present in the sample (**Table 12.2**). Although metaproteomics is less commonly used compared to metabolomics due to its technical challenges, its potential in microbiome research is substantial. Despite challenges, such as the complexity of protein extraction, identification, and quantification from heterogeneous microbiome samples, technological advancements are increasing the

feasibility and accuracy of metaproteomics (Hickl et al., 2019; Karaduta et al., 2021; Sun et al., 2024; Zhang et al., 2025). As analytical tools improve, this approach will continue to gain prominence, offering critical insights into microbial functionality and its impact on host biology and environmental systems (Lloyd-Price et al., 2019; Messer et al., 2024). Page 250

Table 12.2

Key approaches in metaproteomics: advantages and challenges. Advances in shotgun metaproteomics have enabled broad surveys of microbial proteins, while targeted approaches focus on specific functional proteins of interest. Quantification techniques, including label-free and label-based methods, provide insights into protein abundance. When combined with other omics data metaproteomics supports a systems-level understanding of microbial communities, linking taxonomic composition, gene expression, and metabolic activity.

Approach	Description	Advantages	Challenges
Shotgun metaproteomics	Analyzes all detectable proteins in a sample	Comprehensive overview, including low-abundance proteins	Complex data analysis, computationally intensive
Targeted metaproteomics	Focuses on specific proteins or protein families	Higher sensitivity, reduced data complexity, suitable for hypothesis-driven studies	May miss unexpected or novel proteins
Label-free quantification	Quantifies protein abundance based on peak intensity in mass spectra	Simpler, more cost-effective	Less accurate for low-abundance proteins
Label-based quantification	Uses stable isotopes to label proteins for direct comparison	Higher accuracy and precision	More complex sample preparation and analysis
Data integration	Combining with other omics enables linking taxonomic profiles with expressed proteins and metabolic activity	Provides a systems-level view and a more comprehensive understanding of microbial function	Requires sophisticated bioinformatics tools and expertise

Host omics

Although metagenomics, metatranscriptomics, metabolomics, and metaproteomics offer unique insights into microbial composition, gene expression, and functional activity, respectively, they are often insufficient when examined in isolation (Deek et al., 2024; Lloyd-Price et al., 2019). To fully understand the role of microbial communities in health and disease it is crucial to integrate these microbial profiles with host omics (Kolde et al., 2018). Host omics provides valuable information on how microbial activities interact with host biology, shedding light on host responses at multiple levels, including gene regulation, protein expression, and metabolic changes (Chetty & Blekhman, 2024; Heaver et al., 2018; iHMP Research Network Consortium, 2014; Karaduta et al., 2021; Kolde et al., 2018; Zhou et al., 2019). For example microbial metabolites like SCFAs or BAs influence host immune function, signaling pathways, and disease progression, while host transcriptomic data can reveal downstream effects

on gene expression (Lloyd-Price et al., 2019; Schirmer et al., 2018). Well-phenotyped, longitudinal host omics cohorts are already accelerating microbiome research by leveraging extensive population-level host data, providing a cost-effective avenue for uncovering critical host–microbiome interactions while complementing targeted new data collection efforts (Everett et al., 2021; Nguyen et al., 2020).

In summary integrating multiomics data enables a systems-level understanding of the complex host–microbe interplays that drive biological outcomes. This holistic approach not only identifies microbial signatures linked to disease but also highlights host pathways influenced by microbial activity, offering new opportunities for targeted interventions and personalized therapies. In the following section we highlight computational frameworks and fusion paradigms for integrating microbiome multiomics data to uncover functional dynamics, disease mechanisms, and therapeutic potentials of the microbiome in diverse contexts.

Frameworks for microbiome multiomics integration

In multiomics integration a persistent goal is to determine whether integration leads to better prediction or classification performance than single-omics approaches alone while also providing a deeper mechanistic understanding of biological systems. To achieve this multiomics datasets are typically integrated through diagonal, vertical, or horizontal paradigms (**Fig. 12.1**). To help the reader understand how Page 251 multiomics datasets are linked for integrative analysis, it is important to clarify the concept of a shared anchor. A shared anchor in multiomics integration is a common attribute or reference point that enables meaningful integration in a biologically relevant manner. For example vertical integration involves measuring multiple omics layers from the same set of subjects, providing the most direct framework for integrating multiomics data, as shared subjects serve as a natural anchor. Horizontal integration, on the contrary, links datasets where the same omics features are measured across different groups or studies (Zhang et al., 2025). In this case shared features such as microbial taxa or pathways, rather than subjects, serve as the anchor, enabling comparisons across studies or meta-analyses that identify consistent biological signatures across populations (Ma et al., 2022).

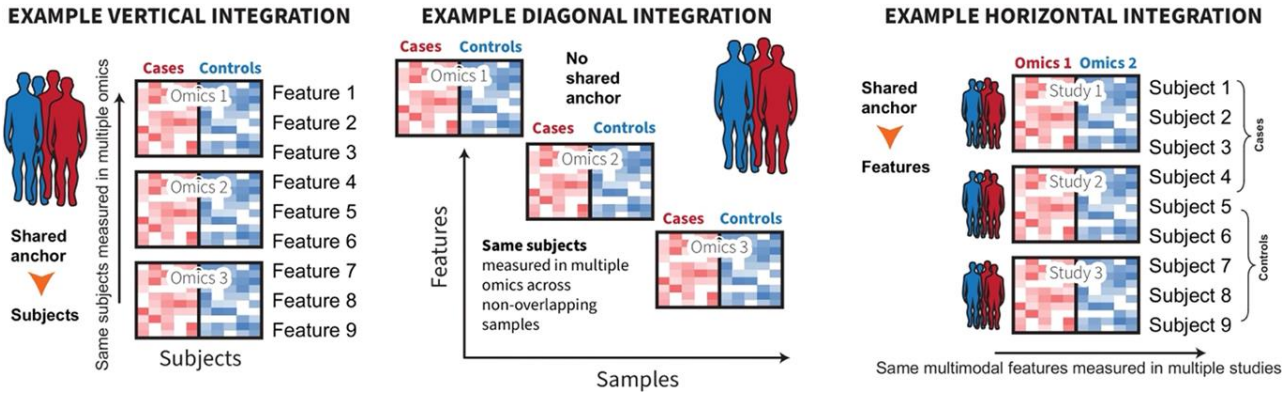


FIGURE 12.1 Frameworks of multimodal data fusion. Vertical integration aligns datasets through a shared anchor of subjects or samples, allowing omics profiles (e.g., microbiome, transcriptomics, and metabolomics) collected from the same individuals to be integrated. We define an anchor as the shared element that aligns datasets across different omics layers, studies, or groups, serving as a reference point for meaningful integration. Note that case/control status serves as an attribute or label describing the subjects or samples rather than functioning as the anchor itself. Horizontal integration uses shared features, such as taxa, genes, or metabolic pathways, to align datasets across different studies or groups. This approach enables researchers to harmonize datasets that may come from distinct populations, environments, or experimental conditions by focusing on common biological or functional units. For example two independent microbiome studies conducted in different geographic regions can be integrated by aligning their data based on overlapping taxa. Similarly metagenomics datasets can be combined by identifying shared genes or pathways that are conserved across the studies. Horizontal integration facilitates cross-study comparisons and meta-analyses, ensuring that insights can be drawn from diverse datasets by leveraging their shared biological features. Diagonal integration lacks a readily available shared anchor, as neither subjects nor features align directly between datasets. Instead computational approaches such as manifold learning are employed to to uncover latent structures that can serve as implicit anchors or alignment references (Argelaguet et al., 2021; Xu & McCord, 2022). These computational approaches allow researchers to infer relationships between datasets that are not directly comparable, enabling the integration of highly heterogeneous data sources and broadening the scope of multimodal integration.

Follow for extended description

In contexts such as multistudy integration or meta-analyses (horizontal integration) factors like consistent protocols, populations, and experimental designs are critical for effective data integration. A key concept in this process is matching, which involves identifying commonalities between datasets to ensure consistency across important Page 252 biological variables. Within the broader concept of matching anchors play a more specific and central role. Anchors are the primary points of linkage between datasets and serve as the backbone of integrative analyses (e.g., subjects in vertical integration and features in horizontal integration). In contrast other matching attributes—such as harmonized sequencing protocols, outcome variables (e.g., case/control status), data preprocessing steps, or demographic variables (e.g., age and sex)—provide context and enhance interpretability but do not function as direct linkages. For example in a microbiome meta-analysis, while consistent sequencing platforms improve data quality, only matching features like shared taxa or pathways act as true anchors, enabling direct integration. Differentiating between these facilitating factors (e.g., age, sex, and case/control status) and true anchors (e.g., subjects or features) is essential for designing robust multiomics integration strategies. Anchors provide the structural foundation for integration, while matching ensures consistency and interpretability, preserving both biological relevance and methodological rigor in integrative studies.

Unlike vertical and horizontal integration, diagonal integration lacks a readily available shared anchor and relies on computational approaches to identify shared latent spaces or axes of variation across nonaligned datasets, enabling cross-modal alignment and harmonization (Argelaguet et al., 2021; Xu & McCord, 2022). For example microbial taxonomic profiles from a metagenomics dataset in one cohort can be linked to metabolomics data from another cohort by projecting both datasets into a shared low-dimensional space or by leveraging external knowledge bases or statistical associations that map microbial pathways to their metabolic products. This type of integration is particularly useful when direct integration is not straightforward. Mallick et al. (2024) recently proposed a two-step approach to achieve diagonal integration in the iHMP: first, aggregating various longitudinally nonaligned omics layers (e.g., metagenomics and metabolomics) at the subject level and then performing vertical integration on the subject-level pseudo-data to develop prediction models for IBD. Diagonal integration thus expands the scope of multiomics studies by allowing connections to be drawn between datasets that lack direct and naturally defined anchors. While challenging it opens new avenues for discovering cross-omics relationships that might otherwise remain hidden, particularly when supplemented with external databases or prior biological knowledge (Xu and McCord, 2022). Page 253

It is also worth noting that multiomics is a specific case of broader multimodal terms such as “multiview,” “multitable,” and “multimodal,” which are often used interchangeably to describe datasets and analyses that integrate multiple perspectives or types of data. The overarching goal of these approaches is to combine data from heterogeneous sources into a unified representation that provides a comprehensive view of the underlying system. Multiomics datasets can be defined at both the biospecimen level and the modality level. At the biospecimen level data may come from different samples, such as stool and tissue from the same individuals, integrating stool metagenomics with tissue transcriptomics for cross-tissue analyses. Alternatively stool samples may be analyzed across multiple omics layers, such as metagenomics, metatranscriptomics, and metabolomics. These are considered distinct omics layers because they provide different but complementary insights by measuring distinct entities, thereby enabling cross-ontological integration. In both cases these layers involve fundamentally different types of biological data and span multiple “omics” domains. In contrast ASVs, MetaPhlAn outputs, or phylogenies derived from assembly-based approaches using the same stool samples are not considered separate omics layers but rather alternative manifestations of microbial taxonomic data. Blindly integrating these manifestations may not be beneficial from an integrative analysis standpoint due to their high degree of overlap, although differences in their resolution may still provide complementary insights. These examples highlight how multiomics datasets can vary in structure and scope, ranging from integration across distinct biospecimens to the harmonization of alternative manifestations of the same data type. Understanding these levels of abstraction is crucial for designing effective multiomics studies and ensuring context-aware integration.

Effectively combining these multiomics datasets requires the application of fusion strategies—early, intermediate, or late fusion—which determine when and how integration occurs (**Fig. 12.2**). Early fusion combines multiomics data by concatenating all features into a single dataset, enabling models to jointly analyze patterns across modalities. While this approach captures comprehensive interactions, it can lead to high dimensionality and noise, potentially overshadowing biologically relevant signals (Ding et al., 2022; Ghaemi et al., 2019). Intermediate fusion integrates multiomics data at an abstract representation level, where features or patterns are extracted and condensed into a manageable format (Ding et al., 2022). For instance dimensionality reduction Page 254 techniques can be applied to individual omics datasets to create latent embeddings, which are then combined into a unified representation for further analysis (Anceschi et al., 2024; Samorodnitsky et al., 2024). This approach captures both shared cross-omics information and modality-specific details, offering a balance between preserving complexity and improving interpretability. Late fusion, in contrast, processes each omics layer independently through separate models and integrates their outputs to make final predictions (Ding et al., 2022; Mallick et al., 2024). This approach preserves the unique characteristics of each omics layer but may miss synergistic cross-omics relationships that are critical for understanding complex biological systems.

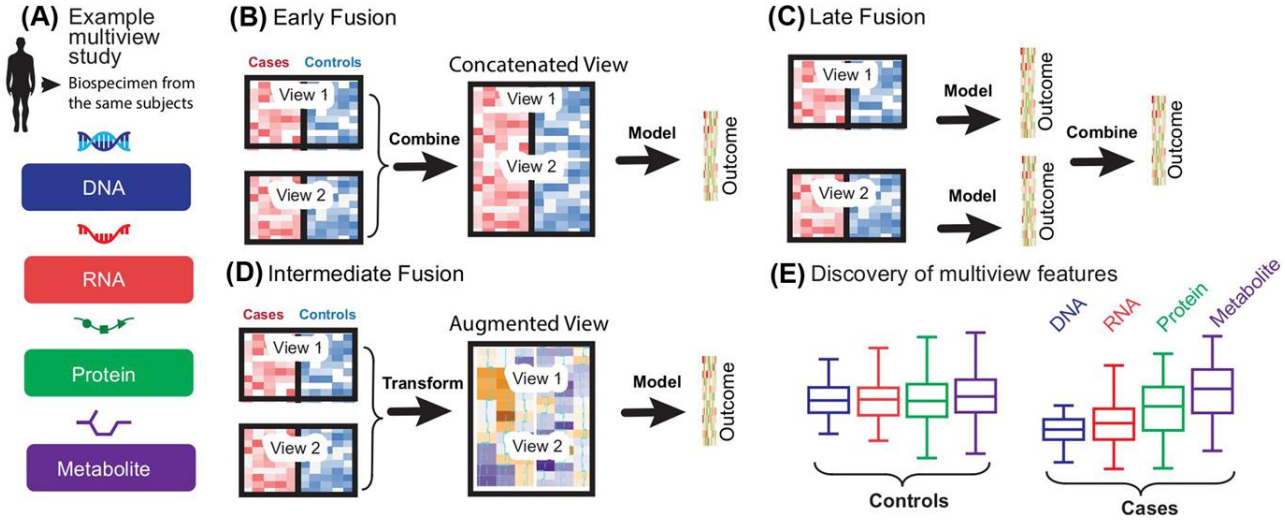


FIGURE 12.2 Three paradigms of multimodal data fusion. A. A hypothetical example of a microbiome multiview study consisting of four modalities collected from the same subjects: DNA, RNA, protein, and metabolite. B. Early fusion concatenates views into a single representation before modeling. C. Late fusion models each view independently and combines the results at a later stage. D. Intermediate fusion transforms views into a shared, intermediate reconstructed view that captures inter-view dependencies prior to modeling. E. Multimodal integration enables the discovery of actionable biological features across modalities, providing a comprehensive understanding of the underlying biological systems. Panel E illustrates the putative outcome of a multiview analysis and demonstrates how shared and unique information across modalities can be harmonized to deliver a more complete understanding of the system under study. As described above multiview analysis often extends beyond simple concatenation (early fusion). While concatenation aligns data, it does not inherently capture interdependencies. Methods such as intermediate address this limitation by modeling both shared and unique contributions across views. These approaches ensure that the analysis is both biologically and statistically meaningful, leveraging interactions between modalities to extract complementary insights.

Follow for extended description

No clear consensus exists on the best approach among early, intermediate, and late fusion strategies (Anceschi et al., 2024; Ding et al., 2022; Mallick et al., 2024; Samorodnitsky et al., 2024). Each method has its strengths and limitations: Late fusion offers computational efficiency and preserves modality-specific signals but risks overlooking synergistic relationships across omics layers; early fusion enables comprehensive interaction modeling but may introduce noise Page 255 and overfitting due to high dimensionality; and intermediate fusion balances shared and unique contributions but requires careful tuning to avoid loss of resolution. The optimal choice depends heavily on the study's goals, data characteristics, and computational resources. To advance the field more extensive and systematic comparative benchmarking across diverse datasets and applications is urgently needed (Anceschi et al., 2024; Ding et al., 2022; Mallick et al., 2024; Samorodnitsky et al., 2024). Such efforts would provide clearer guidance on selecting the most effective fusion strategies for specific research contexts.

The analysis of integrated multiomics datasets can be further categorized into supervised and unsupervised approaches, depending on the use of labeled data in the analysis. Unsupervised approaches, such as similarity network fusion [SNF (Pai & Bader, 2018; Wang et al., 2014)] and multiomics factor analysis [MOFA (Argelaguet, 2018; Argelaguet et al., 2020)], are examples of unsupervised multiomics approaches (**Table 12.3**). These methods do not rely on predefined labels (such as a "case/control" column in the metadata table) but instead aim to uncover hidden patterns, relationships, or clusters within the data, such as novel disease subtypes or molecular signatures. The Human Microbiome Project [HMP (Huttenhower et al., 2012; Lloyd-Price et al., 2017)], which primarily included

healthy samples without explicit disease or case/control labels and lacked a traditional control group, exemplifies unlabeled data. This distinction is important when considering analysis approaches: Supervised methods are particularly effective for prediction tasks tied to specific outcomes or classifications, whereas unsupervised approaches excel particularly well in exploratory settings where patterns are unknown.

Table 12.3

A selected summary of unsupervised multiomics integration methods for multiview clustering and dime reduction.

Method	Short description	Implementation
bayesCC (Lock and Dunson, 2013)	Finite Dirichlet mixture models; supports both overall clustering and a clustering specific to each view	R package: **https://github.com/ttriche/bayesCC**
JIVE (Lock et al., 2013)	Joint/individual variation decomposition with low-rank approximation	R package: **https://cran.r-project.org/web/packages/r.jive/**
iCluster (Shen et al., 2009; Shen et al., 2012), ClusterBayes (Mo et al., 2018)	Gaussian latent variable models with both Bayesian and non-Bayesian algorithms	R package: **https://www.bioconductor.org/packages/devel/bioc/html/i**
Similarity network fusion (SNF) (Wang et al., 2014)	Iteratively fuses networks based on similarity between patients	R package: **https://cran.r-project.org/web/packages/SNFtool/**
NetICS (Dimitrakopoulos et al., 2018)	Ranks genes using per-sample bidirectional network diffusion; integrates sample-specific lists into an overall list	MATLAB package: **https://github.com/cbg-ethz/netics**

Method	Short description	Implementation
Multiomics factor analysis (MOFA) (Argelaguet, 2018), MOFA2 Ricard (Argelaguet et al., 2020)	Provides interpretable low-dimensional representation in terms of latent factors	R package: **https://biofam.github.io/MOFA/**; **https://biofam.gith**
IntNMF (Chalise et al., 2017)	Integrative clustering based on nonnegative matrix factorization	R package: **https://cran.r-project.org/web/packages/IntNMF/**
RGCCA (Tenenhaus et al., 2014)	Finds linear combinations of view variables that explain their correlation structure well; optimized by a monotonically convergent algorithm	R package: **https://cran.r-project.org/web/packages/RGCCA/**

Among unsupervised approaches SNF constructs similarity networks for each omics layer and iteratively fuses them to uncover hidden relationships. MOFA identifies the major axes of variation across omics datasets, revealing underlying drivers of biological heterogeneity without prior assumption. Multiview clustering (MVC) is another unsupervised approach designed to identify clusters or subgroups within multiomics datasets by leveraging information from multiple data modalities or views. Unlike traditional clustering methods MVC accounts for the unique contributions of each data modality while uncovering shared patterns or relationships across views. This makes MVC particularly useful for integrating heterogeneous data and exploring both modality-specific and cross-modal biological signals (Chalise et al., 2017; Dimitrakopoulos et al., 2018; Lock & Dunson, 2013; Tenenhaus et al., 2014). Page 256 Dimensionality reduction methods, in contrast, focus on simplifying high-dimensional datasets by projecting them onto a lower-dimensional space while preserving their essential structure and variability. This approach is vital in multiomics studies, as it reduces computational complexity, mitigates noise, and highlights key features or factors driving biological variation. Techniques such as principal component analysis (PCA), MOFA (Argelaguet, 2018; Argelaguet et al., 2020), and Joint And Individual Variation Explained [JIVE (Lock et al., 2013)] allow researchers to uncover hidden structures, associations, and sources of variation within complex multiomics datasets. Page 257

On the contrary supervised approaches rely on labeled data to predict outcomes or identify discriminative features linked to phenotypic variation. Labeled data refer to datasets where specific outcomes or categories (e.g., healthy vs disease, case vs control, and responder vs nonresponder) are known and explicitly provided. These labels serve as the foundation for training models to predict or classify new data. Methods such as *IntegratedLearner* (Mallick et al., 2024) and *Cooperative Learning* (Ding et al., 2022) are designed for this purpose (**Table 12.4**). By carefully selecting the appropriate fusion strategy and analysis approach researchers can harness the power of multiomics integration to

uncover novel biological insights and improve performance. However best Page 258 practices and centralized resources for microbiome multiomics integration remain in their early stages, leaving room for further development and standardization in the field.

Table 12.4

A selected summary of supervised microbiome multiomics integration methods.

Method	Fusion paradigm	Short description	Implementation
IntegratedLearner (Mallick et al., 2024)	Early, Late	Ensemble learning using Bayesian additive regression trees; supports other machine learning models as base- and meta-learners	R package: **https://github.com/himelmallick/IntegratedLea**
DIABLO (Singh et al., 2019)	Early	Sparse generalized canonical correlation analysis	R package: **http://mixomics.org/**
sPLS-DA (Lê Cao et al., 2011)	Early	Sparse generalized canonical correlation analysis	R package: **http://mixomics.org/**
Cooperative learning (Ding et al., 2022)	Intermediate	Collaborative regression using agreement penalty and LASSO regularization	R package: **https://cran.r-project.org/web/packages/multiview/**
MOGLAM (Ouyang et al., 2023)	Early	Graph learning with attention mechanism	Python package: **https://github.com/Ouyang-Dong/MOGLAM**
Franzosa et al. (2019)	Early	Concatenated random forest	R package: **https://github.com/himelmallick/IntegratedLea**

Method	Fusion paradigm	Short description	Implementation
Stelzer et al. (2021)	Late	Ensemble learning using LASSO	R package: **https://github.com/himelmallick/IntegratedLea**
Ghaemi et al. (2019)	Late	Ensemble learning using elastic net	R package: **https://github.com/himelmallick/IntegratedLea**
BSFP (Samorodnitsky et al.,2024)	Intermediate	Bayesian joint additive factor models	R package: **https://github.com/sarahsamorodnitsky/BSFP**
JAFAR (Anceschi et al., 2024)	Intermediate	Bayesian joint additive factor models	R package: **https://github.com/niccoloanceschi/jafar**

Success stories in microbiome multiomics

Multiomics approaches have significantly advanced our understanding of complex diseases by integrating diverse datasets to reveal intricate relationships between the microbiome, host biology, and disease mechanisms. While it is beyond the scope of this section to provide a fully comprehensive review, several notable case studies across various fields demonstrate the transformative power of multiomics in uncovering biological discoveries not achievable with single-omics approaches. For example Franzosa et al. (2019) found significant alterations in metabolomic and microbiome profiles in individuals with IBD, including disruptions in lipid metabolism and bile acid transformation linked to gut inflammation. Lloyd-Price et al. (2019) observed functional dysbiosis in the IBD microbiome, marked by shifts to facultative anaerobes, disrupted metabolites, and associations with disease activity. Under the iHMP umbrella another study found distinct immune and microbiome responses to respiratory viral infections in insulin-resistant versus insulin-sensitive individuals, suggesting a link between insulin resistance and microbiome variability (Zhou et al., 2019).

Beyond iHMP, in cancer research, multiomics studies have provided critical insights into colorectal cancer (CRC), gastric cancer, and breast cancer. For CRC Zou et al. (2024) combined microbiome and host transcriptomics, identifying 22 microbial species associated with key somatic mutations. They showed that *Fusobacterium nucleatum* positively correlated with host metabolic pathways and influenced tumor immunity by modulating tumor necrosis factor superfamily, member 9 (TNFSF9) expression. Liu et al. (2023) demonstrated that the intratumor microbiome composition affects CRC survival outcomes, particularly immune checkpoint blockade (ICB) therapy response, where *Clostridium* prevalence correlated with ICB resistance. In gastric cancer Park et al. (2022) highlighted microbial shifts across stages, such as *Helicobacteraceae* driving inflammation and *Lachnospiraceae* correlating with cell cycle dysregulation. Cheng et al. (2024) revealed microbial and metabolic markers, such as *Porphyromonadaceae* and arginine, associated with resistance to PD-1/PD-L1 immunotherapy in Page 259 microsatellite-instability-high gastrointestinal cancers. Finally Li et al. (2024)'s breast cancer study uncovered how microbes like *Lachnoclostridium* and *Acidibacillus* influence tumor immunity, prognosis, and immunotherapy responses.

In neurological disorders multiomics has shed light on the gut–brain axis (Dinan & Cryan, 2017). Pereira et al. (2022) integrated 16 S rRNA sequencing, metabolomics and clinical data to differentiate Parkinson's disease (PD) patients from controls, revealing that microbial dysbiosis disrupts lipid and energy metabolism. Similarly Cantoni et al. (2022) explored multiple sclerosis, identifying distinct gut microbiota profiles (e.g., reduced *Faecalibacterium* and

Prevotella), immune imbalances, and elevated lipid metabolites, linking dietary habits to disease progression. Multiomics has also provided insights into autism spectrum disorder (ASD). Morton et al. (2023) conducted a multistudy analysis (horizontal integration) to link microbiome changes in children with ASD to brain gene expression, particularly amino acid and carbohydrate pathways, highlighting dietary influences on metabolic function. Researchers at Arizona State University found that microbiota transplant therapy (MTT) significantly improved gastrointestinal and ASD-related symptoms in children with ASD, while shifting their multiomics profiles to resemble those of typically developing peers, highlighting the potential of multiomics as a biomarker for MTT treatment efficacy (Adams et al., 2019; Kang, 2017; Kang et al., 2019, 2020).

In infectious diseases multiomics approaches have revealed the role of gut microbiome dysbiosis in recurrent urinary tract infections (rUTIs). Worby et al. (2022) showed that women with rUTIs exhibit reduced butyrate-producing bacteria and distinct immune responses, emphasizing the gut–bladder axis role in infection susceptibility. Collectively these case studies highlight how multiomics approaches have facilitated discoveries across diverse diseases by exploiting complex cross-omics interactions. Most importantly all these studies successfully achieved biological insights that would not have been possible using single-omics approaches alone.

Conclusions and future directions

The success stories highlighted in the previous section underscore the immense potential of multiomics to uncover the functional complexities of microbial ecosystems and their profound influence on health and Page 260 disease. By linking microbial metabolites like SCFAs to host metabolic pathways or integrating metagenomics, metabolomics, and host transcriptomics to identify novel biomarkers, these case studies demonstrate how multiomics can move us beyond correlations to reveal potentially causal relationships and mechanistic insights into host–microbe interactions. Going forward, to fully realize this immense potential and translate multiomics insights into actionable outcomes, several key challenges must be addressed.

First the dynamic nature of the microbiome poses a significant challenge to replicability (Schloss, 2018). Microbial communities are in constant flux, shifting in response to diet, environment, and circadian rhythms (David et al., 2014; Pearson et al., 2020). Capturing these changes requires standardized, time-stamped sampling, and longitudinal studies that can separate meaningful biological patterns from background noise (Allaband et al., 2024). Achieving strain-level resolution is another critical hurdle. While species-level analyses have revealed broad patterns, the true functional diversity often lies at the strain level (Ghazi et al., 2022; Yan et al., 2020). Advances such as long-read sequencing are now enabling researchers to resolve these differences with greater accuracy, particularly when paired with metabolomics and host omics to pinpoint which strains drive specific outcomes (Agustinho et al., 2024; Kim et al., 2024).

Diet introduces further complexity, as its effects on the microbiome vary widely across populations due to cultural and regional differences (David et al., 2014; Johnson et al., 2020; Serrano et al., 2019; Singh, 2017; Wilson et al., 2020). Integrating detailed dietary metadata into multiomics studies and building globally representative biobanks will ensure that findings are both equitable and widely applicable, supporting personalized interventions across diverse groups (Ma et al., 2017; Ryan et al., 2021). Technologies like spatial sequencing add yet another dimension by revealing the localization of microbial communities within tissues or environments, uncovering niche-specific interactions that drive functional outcomes (Saarenpää et al., 2024; Sheth et al., 2019). When integrated with other omics layers these tools can provide a richer, more comprehensive view of host–microbe interplays.

Power calculation remains an essential yet often overlooked aspect of multiomics study design (Ma et al., 2021). Given the complexity and high dimensionality of these datasets ensuring adequate statistical power is critical for detecting meaningful biological signals and avoiding false positives or negatives. In addition to the routine sample Page 261 size and effect size requirements typical of single-omics studies multiomics power calculations must also account for the expected variability across omics layers (Hasin et al., 2017; Tarazona et al., 2020, 2021). Incorporating rigorous power analysis into multiomics studies not only enhances replicability but also optimizes resource allocation, particularly in large-scale projects involving diverse modalities (). While simulators for microbiome

multiomics data are still in their early stages, ongoing advancements in this area are paving the way for more robust and tailored power analyses that can address the unique characteristics of these datasets, combining lessons learned from earlier microbiome epidemiology with challenges unique to microbiome multiomics (Chung & Kang, 2019; Ma et al., 2021; Mallick et al., 2017; Monzó et al., 2025).

From the computational side, as datasets grow larger and more complex, scalability remains a pressing concern. Multimodal tools like *MultiAssayExperiment* (Ramos et al., 2017) that facilitate the integration of diverse datasets in a unified manner and *IntegratedLearner* (Mallick et al., 2024) that enables uncertainty quantification must continue to be refined to address the challenges of increasingly common strain-resolved, multimodal epidemiology studies of microbial communities. The promise of large language models, foundation models, causal discovery, and generative AI further enhances this landscape as they offer the ability to identify hidden patterns across complex, multiomics datasets as well as predict functional outputs and generate synthetic profiles, which can be particularly impactful in resource-limited settings (Corander et al., 2022; Cosma et al., 2024; Gerber, 2024; Hannigan et al., 2019; Karkera et al., 2023; Wu et al., 2024; Yan et al., 2025). This parallels early predictive tools such as *MelonnPan* (Mallick et al., 2019) and *PICRUSt* (Douglas et al., 2020), but multimodal integration can significantly enhance the capabilities of these tools moving beyond static prediction to real-time hypothesis generation toward health applications of microbiome research.

Finally the success of multiomics research depends on ensuring ethical and equitable practices (Oduaran & Bhatt, 2022). This includes transparent frameworks for data privacy, participant consent, and global accessibility to tools and findings. Training the next generation of researchers in integrative, interdisciplinary approaches will also be essential for driving innovation and fostering collaboration across fields. Much like how genome-wide association studies (GWAS) revolutionized human genetics, microbiome multiomics has the Page 262 potential to redefine how we understand the microbiome. By addressing these challenges—standardizing protocols, deep phenotyping, achieving strain-level precision, improving computational methods, and expanding global inclusivity—multiomics research is poised to transform human health, ecological sustainability, and personalized medicine in the coming years.

References

Abu-Ali et al., 2018 Abu-Ali GS, Mehta RS, Lloyd-Price J, et al. Metatranscriptome of human faecal microbial communities in a cohort of adult men. *Nature Microbiology*. 2018;3(3):356–366 **https://doi.org/10.1038/s41564-017-0084-4**.

Adams et al., 2019 Adams JB, Borody TJ, Kang DW, Khoruts A, Krajmalnik-Brown R, Sadowsky MJ. Microbiota transplant therapy and autism: Lessons for the clinic. *Expert Review of Gastroenterology and Hepatology*. 2019;13(11):1033–1037 **https://doi.org/10.1080/17474124.2019.1687293**.

Agustinho et al., 2024 Agustinho DP, Fu Y, Menon VK, Metcalf GA, Treangen TJ, Sedlazeck FJ. Unveiling microbial diversity: Harnessing long-read sequencing technology. *Nature Methods*. 2024;21(6):954–966 **https://doi.org/10.1038/s41592-024-02262-1**.

Ala, 2022 Ala M. Tryptophan metabolites modulate inflammatory bowel disease and colorectal cancer by affecting immune system. *International Reviews of Immunology*. 2022;41(3):326–345 **https://doi.org/10.1080/08830185.2021.1954638**.

Allaband et al., 2024 Allaband C, Lingaraju A, Ramos SF, et al. Time of sample collection is critical for the replicability of microbiome analyses. *Nature Metabolism*. 2024;6(7):1282–1293 **https://doi.org/10.1038/s42255-024-01064-1**.

Anceschi et al., 2024 Anceschi N, Ferrari DB, Dunson H, Mallick H. Bayesian joint additive factor models for multiview learning. *arXiv:2406.00778* 2024; **https://doi.org/10.48550/arXiv.2406.00778**.

Argelaguet et al., 2021 Argelaguet R, Cuomo ASE, Stegle O, Marioni JC. Computational principles and challenges in single-cell data integration. *Nature Biotechnology*. 2021;39(10):1202–1215

https://doi.org/10.1038/s41587-021-00895-7.

Argelaguet et al., 2020 Argelaguet R, Arnol D, Bredikhin D, et al. MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*. 2020;21(1):111 **https://doi.org/10.1186/s13059-020-02015-1**.

Argelaguet et al., 2018 Argelaguet R, Velten B, Arnol D, et al. Multi-omics factor analysis—A framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*. 2018;14:e8124 **https://doi.org/10.15252/msb.20178124**.

Bauermeister et al., 2022 Bauermeister A, Mannochio-Russo H, Costa-Lotufo LV, Jarmusch AK, Dorrestein PC. Mass spectrometry-based metabolomics in microbiome investigations. *Nature Reviews Microbiology*. 2022;20(3):143–160 **https://doi.org/10.1038/s41579-021-00621-9**.

Beale et al., 2016 Beale DJ, Karpe AV, Ahmed W. *Beyond metabolomics: A review of multi-omics-based approaches*. *Microbial metabolomics: Applications in clinical, environmental, and industrial microbiology* Springer International Publishing 2016;289–312 **http://doi.org/10.1007/978-3-319-46326-1**.

Cantoni et al., 2022 Cantoni C, Lin Q, Dorsett Y, et al. Alterations of host-gut microbiome interactions in multiple sclerosis. *eBioMedicine*. 2022;76:103798 **https://doi.org/10.1016/j.ebiom.2021.103798**.

Chalise et al., 2017 Chalise P, Fridley BL, Peddada SD. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. *PLoS One*. 2017;12(5):e0176278 **https://doi.org/10.1371/journal.pone.0176278**.

Chen et al., 2019 Chen MX, Wang SY, Kuo CH, Tsai IL. Metabolome analysis for investigating host-gut microbiota interactions. *Journal of the Formosan Medical Association*. 2019;118:S10–S22 **https://doi.org/10.1016/j.jfma.2018.09.007**.

Cheng et al., 2024 Cheng S, Han Z, Dai D, et al. Multi-omics of the gut microbial ecosystem in patients with microsatellite-instability-high gastrointestinal cancer resistant to immunotherapy. *Cell Reports Medicine*. 2024;5(1):101355 **https://doi.org/10.1016/j.xcrm.2023.101355**.

Chetty and Blekhman, 2024 Chetty A, Blekhman R. Multi-omic approaches for host-microbiome data integration. *Gut Microbes*. 2024;16(1):2297860 **https://doi.org/10.1080/19490976.2023.2297860**.

Chung and Kang, 2019 Chung RH, Kang CY. A multi-omics data simulator for complex disease studies and its application to evaluate multi-omics data analysis methods for disease classification. *GigaScience*. 2019;8(5):giz045 **https://doi.org/10.1093/gigascience/giz045**.

Collins et al., 2023 Collins SL, Stine JG, Bisanz JE, Okafor CD, Patterson AD. Bile acids and the gut microbiota: Metabolic interactions and impacts on disease. *Nature Reviews Microbiology*. 2023;21(4):236–247 **https://doi.org/10.1038/s41579-022-00805-x**.

Corander et al., 2022 Corander J, Hanage WP, Pensar J. Causal discovery for the microbiome. *The Lancet Microbe*. 2022;3(11):e881 **https://doi.org/10.1016/S2666-5247(22)00186-0**.

Cosma et al., 2024 Cosma BM, Pillay S, Calderon-Franco D, Abeel T. Predicted meta-omics: A potential solution to multi-omics data scarcity in microbiome studies. *bioRxiv* 2024; **https://doi.org/10.1101/2024.11.04.621857**.

Danzi et al., 2023 Danzi F, Pacchiana R, Mafficini A, et al. To metabolomics and beyond: A technological portfolio to investigate cancer metabolism. *Signal Transduction and Targeted Therapy*. 2023;8(1):137 **https://doi.org/10.1038/s41392-023-01380-0**.

David et al., 2014 David LA, Maurice CF, Carmody RN, et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature*. 2014;505(7484):559–563 **https://doi.org/10.1038/nature12820**.

Deek et al., 2024 Deek RA, Ma S, Lewis J, Li H. Statistical and computational methods for integrating microbiome, host genomics, and metabolomics data. *eLife*. 2024;13:e88956 **https://doi.org/10.7554/elife.88956**.

Dimitrakopoulos et al., 2018 Dimitrakopoulos C, Hindupur SK, Häfliger L, et al. Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics (Oxford, England)*. 2018;34(14):2441–2448 **https://doi.org/10.1093/bioinformatics/bty148**.

Dinan and Cryan, 2017 Dinan TG, Cryan JF. The microbiome-gut-brain axis in health and disease. *Gastroenterology Clinics of North America*. 2017;46(1):77–89 **https://doi.org/10.1016/j.gtc.2016.09.007**.

Ding et al., 2022 Ding DY, Li S, Narasimhan B, Tibshirani R. Cooperative learning for multiview analysis. *Proceedings of the National Academy of Sciences*. 2022;119 **https://doi.org/10.1073/pnas.2202113119**.

Douglas et al., 2020 Douglas GM, Maffei VJ, Zaneveld JR, et al. PICRUSt2 for prediction of metagenome functions. *Nature Biotechnology*. 2020;38(6):685–688 **https://doi.org/10.1038/s41587-020-0548-6**.

Everett et al., 2021 Everett C, Li C, Wilkinson JE, et al. Overview of the microbiome among nurses study (micro-N) as an example of prospective characterization of the microbiome within cohort studies. *Nature Protocols*. 2021;16(6):2724–2731 **https://doi.org/10.1038/s41596-021-00519-z**.

Fettweis et al., 2019 Fettweis JM, Serrano MG, Brooks JP, et al.... The vaginal microbiome and preterm birth. *Nature Medicine*. 2019;25(6):1012–1021 **https://doi.org/10.1038/s41591-019-0450-2**.

Fleishman and Kumar, 2024 Fleishman JS, Kumar S. Bile acid metabolism and signaling in health and disease: Molecular mechanisms and therapeutic targets. *Signal Transduction and Targeted Therapy*. 2024;9(1):97 **https://doi.org/10.1038/s41392-024-01811-6**.

Franzosa et al., 2019 Franzosa EA, Sirota-Madi A, Avila-Pacheco J, et al.... Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature Microbiology*. 2019;4(2):293–305 **https://doi.org/10.1038/s41564-018-0306-4**.

Gallagher et al., 2021 Gallagher K, Catesson A, Griffin JL, Holmes E, Williams HRT. Metabolomic analysis in inflammatory bowel disease: A systematic review. *Journal of Crohn's and Colitis*. 2021;15(5):813–826 **https://doi.org/10.1093/ecco-jcc/jjaa227**.

Gerber, 2024 Gerber GK. AI in microbiome research: Where have we been, where are we going?. *Cell Host and Microbe*. 2024;32(8):1230–1234 **https://doi.org/10.1016/j.chom.2024.07.021**.

Ghaemi et al., 2019 Ghaemi MS, DiGiulio DB, Contrepois K, et al.... Multiomics modeling of the immunome, transcriptome, microbiome, proteome and metabolome adaptations during human pregnancy. *Bioinformatics (Oxford, England)*. 2019;35(1):95–103 **https://doi.org/10.1093/bioinformatics/bty537**.

Ghazi et al., 2022 Ghazi AR, Münch PC, Chen D, Jensen J, Huttenhower C. Strain identification and quantitative analysis in microbial communities. *Journal of Molecular Biology*. 2022;434(15):167582 **https://doi.org/10.1016/j.jmb.2022.167582**.

Gilbert et al., 2018 Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. *Nature Medicine*. 2018;24(4):392–400 **https://doi.org/10.1038/nm.4517**.

Hannigan et al., 2019 Hannigan GD, Prihoda D, Palicka A, et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Research*. 2019;47(18):e110 **https://doi.org/10.1093/nar/gkz654**.

Hasin et al., 2017 Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biology*. 2017;18(1):1–15 **https://doi.org/10.1186/s13059-017-1215-1**.

Heaver et al., 2018 Heaver SL, Johnson EL, Ley RE. Sphingolipids in host–microbial interactions. *Current Opinion in Microbiology*. 2018;43:92–99 **https://doi.org/10.1016/j.mib.2017.12.011**.

Hickl et al., 2019 Hickl O, Heintz-Buschart A, Trautwein-Schult A, et al. Sample preservation and storage significantly impact taxonomic and functional profiles in metaproteomics studies of the human gut

microbiome. *Microorganisms*. 2019;7(9):367 **https://doi.org/10.3390/microorganisms7090367**.

Huttenhower et al., 2012 Huttenhower C, Gevers D, Knight R, et al. Structure, function and diversity of the healthy human microbiome. *Nature*. 2012;486(7402):207–214 **https://doi.org/10.1038/nature11234**.

iHMP Research Network Consortium, 2014 iHMP Research Network Consortium. The Integrative Human Microbiome Project: Dynamic Analysis of Microbiome-Host Omics Profiles during Periods of Human Health and Disease. *Cell Host & Microbe*. 2014;16(3):276–289 **https://doi.org/10.1016/j.chom.2014.08.014**.

Johnson et al., 2020 Johnson AJ, Zheng JJ, Kang JW, Saboe A, Knights D, Zivkovic AM. A guide to diet-microbiome study design. *Frontiers in Nutrition*. 2020;7:79 **https://doi.org/10.3389/fnut.2020.00079**.

Kang et al., 2019 Kang DW, Adams JB, Coleman DM, et al. Long-term benefit of microbiota transfer therapy on autism symptoms and gut microbiota. *Scientific Reports*. 2019;9(1):5821 **https://doi.org/10.1038/s41598-019-42183-0**.

Kang et al., 2020 Kang DW, Adams JB, Vargason T, Santiago M, Hahn J, Krajmalnik-Brown R. Distinct fecal and plasma metabolites in children with autism spectrum disorders and their modulation after microbiota transfer therapy. *mSphere*. 2020;5:5 **https://doi.org/10.1128/mSphere.00314-20**.

Kang et al., 2017 Kang DW, Adams JB, Gregory AC, et al. Microbiota transfer therapy alters gut ecosystem and improves gastrointestinal and autism symptoms: An open-label study. *Microbiome*. 2017;5(1):10 **https://doi.org/10.1186/s40168-016-0225-7**.

Karaduta et al., 2021 Karaduta, Dvanajscak, Zybailov. Metaproteomics—An advantageous option in studies of host-microbiota interaction. *Microorganisms*. 2021;9(5):980 **https://doi.org/10.3390/microorganisms9050980**.

Karkera et al., 2023 Karkera N, Acharya S, Palaniappan SK. Leveraging pre-trained language models for mining microbiome-disease relationships. *BMC Bioinformatics*. 2023;24(1):290 **https://doi.org/10.1186/s12859-023-05411-z**.

Karu et al., 2018 Karu N, Deng L, Slae M, et al. A review on human fecal metabolomics: Methods, applications and the human fecal metabolome database. *Analytica Chimica Acta*. 2018;1030:1–24 **https://doi.org/10.1016/j.aca.2018.05.031**.

Kim et al., 2024 Kim C, Pongpanich M, Porntaveetus T. Unraveling metagenomics through long-read sequencing: A comprehensive review. *Journal of Translational Medicine*. 2024;22(1):111 **https://doi.org/10.1186/s12967-024-04917-1**.

Kolde et al., 2018 Kolde R, Franzosa EA, Rahnavard G, et al. Host genetic variation and its microbiome interactions within the Human Microbiome Project. *Genome Medicine*. 2018;10(1):1–13 **https://doi.org/10.1186/s13073-018-0515-8**.

Krautkramer et al., 2021 Krautkramer KA, Fan J, Bäckhed F. Gut microbial metabolites as multi-kingdom intermediates. *Nature Reviews Microbiology*. 2021;19(2):77–94 **https://doi.org/10.1038/s41579-020-0438-4**.

Larabi et al., 2023 Larabi AB, Masson HLP, Bäumler AJ. Bile acids as modulators of gut microbiota composition and function. *Gut Microbes*. 2023;15(1):2172671 **https://doi.org/10.1080/19490976.2023.2172671**.

Lavelle and Sokol, 2018 Lavelle A, Sokol H. Beyond metagenomics, metatranscriptomics illuminates microbiome functionality in IBD. *Nature Reviews Gastroenterology & Hepatology*. 2018;15(4):193–194 **https://doi.org/10.1038/nrgastro.2018.15**.

Lê Cao et al., 2011 Lê Cao KA, Boitard S, Besse P. Sparse PLS discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*. 2011;12:1–17 **https://doi.org/10.1186/1471-2105-12-253**.

Li et al., 2024 Li J, Zhang Y, Cai Y, et al. Multi-omics analysis elucidates the relationship between intratumor microbiome and host immune heterogeneity in breast cancer. *Microbiology Spectrum*. 2024;12(4):e04104–23 **https://doi.org/10.1128/spectrum.04104-23**.

Liu et al., 2023 Liu Z, Zhang X, Zhang H, et al. Multi-omics analysis reveals intratumor microbes as immunomodulators in colorectal cancer. *Microbiology Spectrum*. 2023;11(2):e05038–22 **https://doi.org/10.1128/spectrum.05038-22**.

Lloyd-Price et al., 2019 Lloyd-Price J, Arze C, Ananthakrishnan AN, et al.... Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*. 2019;569(7758):655–662 **https://doi.org/10.1038/s41586-019-1237-9**.

Lloyd-Price et al., 2017 Lloyd-Price J, Mahurkar A, Rahnavard G, et al. Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature*. 2017;550(7674):61–66 **https://doi.org/10.1038/nature23889**.

Lock and Dunson, 2013 Lock EF, Dunson DB. Bayesian consensus clustering. *Bioinformatics (Oxford, England)*. 2013;29(20):2610–2616 **https://doi.org/10.1093/bioinformatics/btt425**.

Lock et al., 2013 Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Annals of Applied Statistics*. 2013;7(1):523–542 **https://doi.org/10.1214/12-AOAS597**.

Luan et al., 2019 Luan H, Wang X, Cai Z. Mass spectrometry-based metabolomics: Targeting the crosstalk between gut microbiota and brain in neurodegenerative disorders. *Mass Spectrometry Reviews*. 2019;38(1):22–33 **https://doi.org/10.1002/mas.21553**.

Ma et al., 2024 Ma S, Huttenhower C, Janson L. Compositional differential abundance testing: Defining and finding a new type of health-microbiome associations. *bioRxiv*. 2024;2024.06.04.596112 **https://doi.org/10.1101/2024.06.04.596112**.

Ma et al., 2021 Ma S, Ren B, Mallick H, et al. A statistical model for describing and simulating microbial community profiles. *PLoS Computational Biology*. 2021;17(9):e1008913 **https://doi.org/10.1371/journal.pcbi.1008913**.

Ma et al., 2022 Ma S, Shungin D, Mallick H, et al. Population structure discovery in meta-analyzed microbial communities and inflammatory bowel disease using MMUPHin. *Genome Biology*. 2022;23(1):208 **https://doi.org/10.1186/s13059-022-02753-4**.

Ma et al., 2017 Ma Y, Chen H, Lei R, Ren J. Biobanking for human microbiome research: Promise, risks, and ethics. *Asian Bioethics Review*. 2017;9(4):311–324 **https://doi.org/10.1007/s41649-017-0033-9**.

Ma et al., 2024 Ma Z, Zuo T, Frey N, Rangrez AY. A systematic framework for understanding the microbiome in human health and disease: From basic principles to clinical translation. *Signal Transduction and Targeted Therapy*. 2024;9(1):237 **https://doi.org/10.1038/s41392-024-01946-6**.

Mallick et al., 2019 Mallick H, Franzosa EA, Mclver LJ, et al. Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. *Nature Communications*. 2019;10(1):3136 **https://doi.org/10.1038/s41467-019-10927-1**.

Mallick et al., 2017 Mallick H, Ma S, Franzosa EA, Vatanen T, Morgan XC, Huttenhower C. Experimental design and quantitative analysis of microbial community multiomics. *Genome Biology*. 2017;18(1):1–16 **https://doi.org/10.1186/s13059-017-1359-z**.

Mallick et al., 2024 Mallick H, Porwal A, Saha S, Basak P, Svetnik V, Paul E. An integrated Bayesian framework for multi-omics prediction and classification. *Statistics in Medicine*. 2024;43(5):983–1002 **https://doi.org/10.1002/sim.9953**.

Mallick et al., 2021 Mallick H, Rahnavard A, McIver LJ, et al. Multivariable association discovery in population-scale meta-omics studies. *PLoS Computational Biology*. 2021;17(11):e1009442 **https://doi.org/10.1371/journal.pcbi.1009442**.

Mann et al., 2024 Mann ER, Lam YK, Uhlig HH. Short-chain fatty acids: Linking diet, the microbiome and immunity. *Nature Reviews Immunology*. 2024;24(8):577–595 **https://doi.org/10.1038/s41577-024-01014-8**.

Mehta et al., 2018 Mehta RS, Abu-Ali GS, Drew DA, et al. Stability of the human faecal microbiome in a cohort of adult men. *Nature Microbiology*. 2018;3(3):347–355 **https://doi.org/10.1038/s41564-017-0096-0**.

Messer et al., 2024 Messer LF, Lee CE, Wattiez R, Matallana-Surget S. Novel functional insights into the microbiome inhabiting marine plastic debris: Critical considerations to counteract the challenges of thin biofilms using multi-omics and comparative metaproteomics. *Microbiome*. 2024;12(1):36 **https://doi.org/10.1186/s40168-024-01751-x**.

Miao et al., 2021 Miao Z, Humphreys BD, McMahon AP, Kim J. Multi-omics integration in the age of million single-cell data. *Nature Reviews Nephrology*. 2021;17(11):710–724 **https://doi.org/10.1038/s41581-021-00463-x**.

Misheva et al., 2021 Misheva M, Ilott NE, McCullagh JSO. Recent advances and future directions in microbiome metabolomics. *Current Opinion in Endocrine and Metabolic Research*. 2021;20:100283 **https://doi.org/10.1016/j.coemr.2021.07.001**.

Mo et al., 2018 Mo Q, Shen R, Guo C, Vannucci M, Chan KS, Hilsenbeck SG. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics (Oxford, England)*. 2018;19(1):71–86 **https://doi.org/10.1093/biostatistics/kxx017**.

Monzó et al., 2025 Monzó C, Martínez-Mira C, Arzalluz-Luque Á, Conesa A, Tarazona S. MOSim: Bulk and single-cell multi-layer regulatory network simulator. *Briefings in Bioinformatics*. 2025;26(2):bbaf110 **https://doi.org/10.1093/bib/bbaf110**.

Morrison and Preston, 2016 Morrison DJ, Preston T. Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. *Gut Microbes*. 2016;7(3):189–200 **https://doi.org/10.1080/19490976.2015.1134082**.

Morton et al., 2023 Morton JT, Jin DM, Mills RH, et al.... Multi-level analysis of the gut–brain axis shows autism spectrum disorder-associated molecular and microbial profiles. *Nature Neuroscience*. 2023;26(7):1208–1217 **https://doi.org/10.1038/s41593-023-01361-0**.

Nguyen et al., 2020 Nguyen LH, Ma W, Wang DD, et al.... Association between sulfur-metabolizing bacterial communities in stool and risk of distal colorectal cancer in men. *Gastroenterology*. 2020;158(5):1313–1325 **https://doi.org/10.1053/j.gastro.2019.12.029**.

Nickols, 2024 Nickols WA. MaAsLin 3: Refining and extending generalized multivariable linear models for meta-omic association discovery. *bioRxiv* 2024; **https://doi.org/10.1101/2024.12.13.628459**.

Oduaran and Bhatt, 2022 Oduaran OH, Bhatt AS. Equitable partnerships and the path to inclusive, innovative and impactful human microbiome research. *Nature Reviews Gastroenterology and Hepatology*. 2022;19(11):683–684 **https://doi.org/10.1038/s41575-022-00689-5**.

Ouyang et al., 2023 Ouyang D, Liang Y, Li L, et al. Integration of multi-omics data using adaptive graph learning and attention mechanism for patient classification and biomarker identification. *Computers in Biology and Medicine*. 2023;164:107303 **https://doi.org/10.1016/j.compbiomed.2023.107303**.

Pai and Bader, 2018 Pai S, Bader GD. Patient similarity networks for precision medicine. *Journal of Molecular Biology*. 2018;430(18):2924–2938 **https://doi.org/10.1016/j.jmb.2018.05.037**.

Park et al., 2022 Park CH, Hong C, Lee AR, Sung J, Hwang TH. Multi-omics reveals microbiome, host gene expression, and immune landscape in gastric carcinogenesis. *iScience*. 2022;25(3):103956 **https://doi.org/10.1016/j.isci.2022.103956**.

Pearson et al., 2020 Pearson JA, Wong FS, Wen L. Crosstalk between circadian rhythms and the microbiota. *Immunology*. 2020;161(4):278–290 **https://doi.org/10.1111/imm.13278**.

Pereira et al., 2022 Pereira PAB, Trivedi DK, Silverman J, et al. Multiomics implicate gut microbiota in altered lipid and energy metabolism in Parkinson's disease. *npj Parkinson's Disease*. 2022;8(1):39 **https://doi.org/10.1038/s41531-022-00300-3**.

Proctor et al., 2019 Proctor LM, Creasy HH, Fettweis JM, et al. The Integrative Human Microbiome Project. *Nature*. 2019;569(7758):641–648 **https://doi.org/10.1038/s41586-019-1238-8**.

Ramos et al., 2017 Ramos M, Schiffer L, Re A, et al. Software for the integration of multiomics experiments in bioconductor. *Cancer Research*. 2017;77(21):e39 **https://doi.org/10.1158/0008-5472.can-17-0344**.

Ratajczak et al., 2019 Ratajczak W, Rył A, Mizerski A, Walczakiewicz K, Sipak O, Laszczyńska M. Immunomodulatory potential of gut microbiome-derived shortchain fatty acids (SCFAs). *Acta Biochimica Polonica*. 2019;66(1):1–12 **https://doi.org/10.18388/abp.2018_2648**.

Roager and Licht, 2018 Roager HM, Licht TR. Microbial tryptophan catabolites in health and disease. *Nature Communications*. 2018;9(1):3294 **https://doi.org/10.1038/s41467-018-05470-4**.

Ryan et al., 2021 Ryan MJ, Schloter M, Berg G, et al.... Development of microbiome biobanks—Challenges and opportunities. *Trends in Microbiology*. 2021;29(2):89–92 **https://doi.org/10.1016/j.tim.2020.06.009**.

Saarenpää et al., 2024 Saarenpää S, Shalev O, Ashkenazy H, et al. Spatial metatranscriptomics resolves host–bacteria–fungi interactomes. *Nature Biotechnology*. 2024;42(9):1384–1393 **https://doi.org/10.1038/s41587-023-01979-2**.

Saha et al., 2016 Saha S, Rajpal DK, Brown JR. Human microbial metabolites as a source of new drugs. *Drug Discovery Today*. 2016;21(4):692–698 **https://doi.org/10.1016/j.drudis.2016.02.009**.

Samorodnitsky et al., 2024 Samorodnitsky S, Wendt CH, Lock EF. Bayesian simultaneous factorization and prediction using multi-omic data. *Computational Statistics and Data Analysis*. 2024;197:107974 **https://doi.org/10.1016/j.csda.2024.107974**.

Schirmer et al., 2018 Schirmer M, Franzosa EA, Lloyd-Price J, et al.... Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. *Nature Microbiology*. 2018;3(3):337–346 **https://doi.org/10.1038/s41564-017-0089-z**.

Schloss, 2018 Schloss PD. Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. *mBio*. 2018;9(3):e00525–18 **https://doi.org/10.1128/mBio.00525-18**.

Serrano et al., 2019 Serrano MG, Parikh HI, Brooks JP, et al.... Racioethnic diversity in the dynamics of the vaginal microbiome during pregnancy. *Nature Medicine*. 2019;25(6):1001–1011 **https://doi.org/10.1038/s41591-019-0465-8**.

Shaffer et al., 2022 Shaffer JP, Nothias LF, Thompson LR, et al.... Standardized multi-omics of Earth's microbiomes reveals microbial and metabolite diversity. *Nature Microbiology*. 2022;7(12):2128–2150 **https://doi.org/10.1038/s41564-022-01266-x**.

Shen et al., 2009 Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics (Oxford, England)*. 2009;25(22):2906–2912 **https://doi.org/10.1093/bioinformatics/btp543**.

Shen et al., 2012 Shen R, Mo Q, Schultz N, et al. Integrative subtype discovery in glioblastoma using iCluster. *PLoS One*. 2012;7(4):e35236 **https://doi.org/10.1371/journal.pone.0035236**.

Sheth et al., 2019 Sheth RU, Li M, Jiang W, Sims PA, Leong KW, Wang HH. Spatial metagenomic characterization of microbial biogeography in the gut. *Nature Biotechnology*. 2019;37(8):877–883 **https://doi.org/10.1038/s41587-019-0183-2**.

Singh et al., 2019 Singh A, Shannon CP, Gautier B, et al. DIABLO: An integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics (Oxford, England)*. 2019;35(17):3055–3062 **https://doi.org/10.1093/bioinformatics/bty1054**.

Singh et al., 2017 Singh RK, Chang HW, Yan D, et al. Influence of diet on the gut microbiome and implications for human health. *Journal of Translational Medicine*. 2017;15:73 **https://doi.org/10.1186/s12967-017-1175-y**.

Smirnov et al., 2016 Smirnov KS, Maier TV, Walker A, et al. Challenges of metabolomics in human gut microbiota research. *International Journal of Medical Microbiology*. 2016;306(5):266–279 **https://doi.org/10.1016/j.ijmm.2016.03.006**.

Stelzer et al., 2021 Stelzer IA, Ghaemi MS, Han X, et al.... Integrated trajectories of the maternal metabolome, proteome, and immunome predict labor onset. *Science Translational Medicine*. 2021;13(592):abd9898 **https://doi.org/10.1126/scitranslmed.abd9898**.

Sun et al., 2024 Sun Z, Ning Z, Figeys D. *The landscape and perspectives of the human gut metaproteomics*. *Molecular & Cellular Proteomics*. 2024;23(5):100763 **https://doi.org/10.1016/j.mcpro.2024.100763** 100763.

Tarazona et al., 2021 Tarazona S, Arzalluz-Luque A, Conesa A. Undisclosed, unmet and neglected challenges in multi-omics studies. *Nature Computational Science*. 2021;1(6):395–402 **https://doi.org/10.1038/s43588-021-00086-z**.

Tarazona et al., 2020 Tarazona S, Balzano-Nogueira L, Gómez-Cabrero D, et al. Harmonization of quality metrics and power calculation in multi-omic studies. *Nature Communications*. 2020;11(1):3092 **https://doi.org/10.1038/s41467-020-16937-8**.

Tenenhaus et al., 2014 Tenenhaus A, Philippe C, Guillemot V, Le Cao K-A, Grill J, Frouin V. Variable selection for generalized canonical correlation analysis. *Biostatistics (Oxford, England)*. 2014;15(3):569–583 **https://doi.org/10.1093/biostatistics/kxu001**.

Wang et al., 2014 Wang B, Mezlini AM, Demir F, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*. 2014;11(3):333–337 **https://doi.org/10.1038/nmeth.2810**.

Wilson et al., 2020 Wilson AS, Koller KR, Ramaboli MC, et al. Diet and the human gut microbiome: An International Review. *Digestive Diseases and Sciences*. 2020;65(3):723–740 **https://doi.org/10.1007/s10620-020-06112-w**.

Worby et al., 2022 Worby CJ, Schreiber HL, Straub TJ, et al.... Longitudinal multi-omics analyses link gut microbiome dysbiosis with recurrent urinary tract infections in women. *Nature Microbiology*. 2022;7(5):630–639 **https://doi.org/10.1038/s41564-022-01107-x**.

Wu et al., 2024 Wu S, Zhou H, Chen D, Lu Y, Li Y, Qiao J. Multi-omic analysis tools for microbial metabolites prediction. *Briefings in Bioinformatics*. 2024;25(4):bbae264 **https://doi.org/10.1093/bib/bbae264**.

Xu and McCord, 2022 Xu Y, McCord RP. Diagonal integration of multimodal single-cell data: Potential pitfalls and paths forward. *Nature Communications*. 2022;13(1):3505 **https://doi.org/10.1038/s41467-022-31104-x**.

Yan et al., 2025 Yan B, Nam Y, Li L, Deek RA, Li H, Ma S. Recent advances in deep learning and language models for studying the microbiome. *Frontiers in Genetics*. 2025;15:1494474 **https://doi.org/10.3389/fgene.2024.1494474**.

Yan et al., 2020 Yan Y, Nguyen LH, Franzosa EA, Huttenhower C. Strain-level epidemiology of microbial communities and the human microbiome. *Genome Medicine*. 2020;12:71 **https://doi.org/10.1186/s13073-020-00765-y**.

Zeki et al., 2020 Zeki, Eylem, Reçber, Kır, Nemutlu. Integration of GC–MS and LC–MS for untargeted metabolomics profiling. *Journal of Pharmaceutical and Biomedical Analysis*. 2020;190:113509 **https://doi.org/10.1016/j.jpba.2020.113509**.

Zhang et al., 2025 Zhang X, Mallick H, Rahnavard A. Meta-analytic microbiome target discovery for immune checkpoint inhibitor response in advanced melanoma. *bioRxiv* 2025; **https://doi.org/10.1101/2025.03.21.644637**.

Zhang et al., 2025 Zhang X, Ning Z, Mayne J, Figeys D. Clinical microbiome analysis by mass spectrometry–based metaproteomics. *Annual Review of Analytical Chemistry*. 2025;18 **https://doi.org/10.1146/annurev-anchem-071124-113819**.

Zhang et al., 2021 Zhang Y, Thompson KN, Huttenhower C, Franzosa EA. Statistical approaches for differential expression analysis in metatranscriptomics. *Bioinformatics (Oxford, England)*. 2021;37(Supplement_1):i34–i41 **https://doi.org/10.1093/bioinformatics/btab327**.

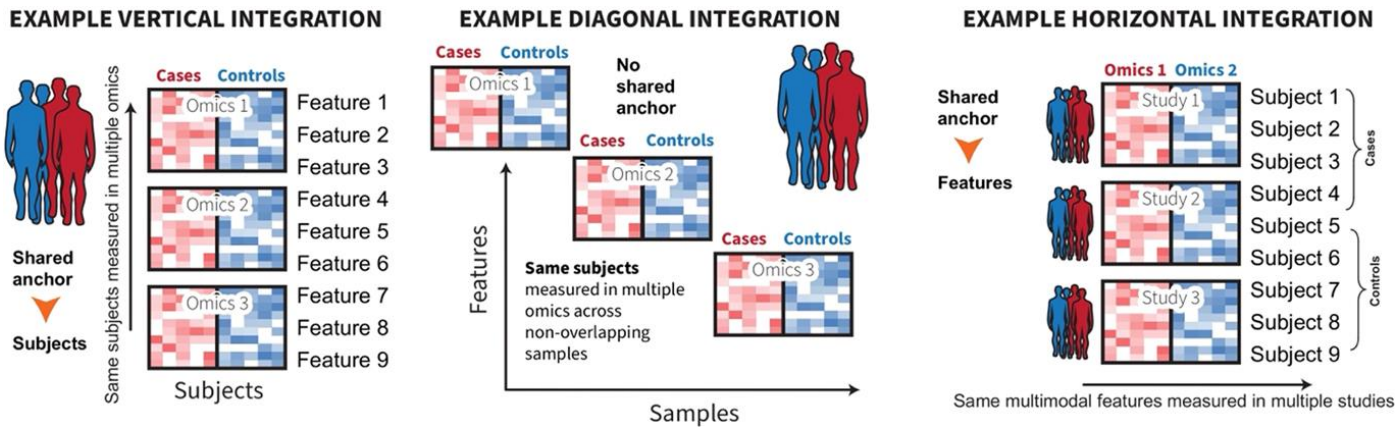
Zhang et al., 2021 Zhang Y, Thompson KN, Branck T, et al. Metatranscriptomics for the human microbiome and microbial community functional profiling. *Annual Review of Biomedical Data Science*. 2021;4(1):279–311 **https://doi.org/10.1146/annurev-biodatasci-031121-103035**.

Zhou et al., 2021 Zhou L, Yu D, Zheng S, Ouyang R, Wang Y, Xu G. Gut microbiota-related metabolome analysis based on chromatography-mass spectrometry. *TrAC Trends in Analytical Chemistry*. 2021;143:116375 **https://doi.org/10.1016/j.trac.2021.116375**.

Zhou et al., 2019 Zhou W, Sailani MR, Contrepois K, et al.... Longitudinal multi-omics of host–microbe dynamics in prediabetes. *Nature*. 2019;569(7758):663–671 **https://doi.org/10.1038/s41586-019-1236-x**.

Zou et al., 2024 Zou S, Yang C, Zhang J, et al. Multi-omic profiling reveals associations between the gut microbiome, host genome and transcriptome in patients with colorectal cancer. *Journal of Translational Medicine*. 2024;22:175 **https://doi.org/10.1186/s12967-024-04984-4**.

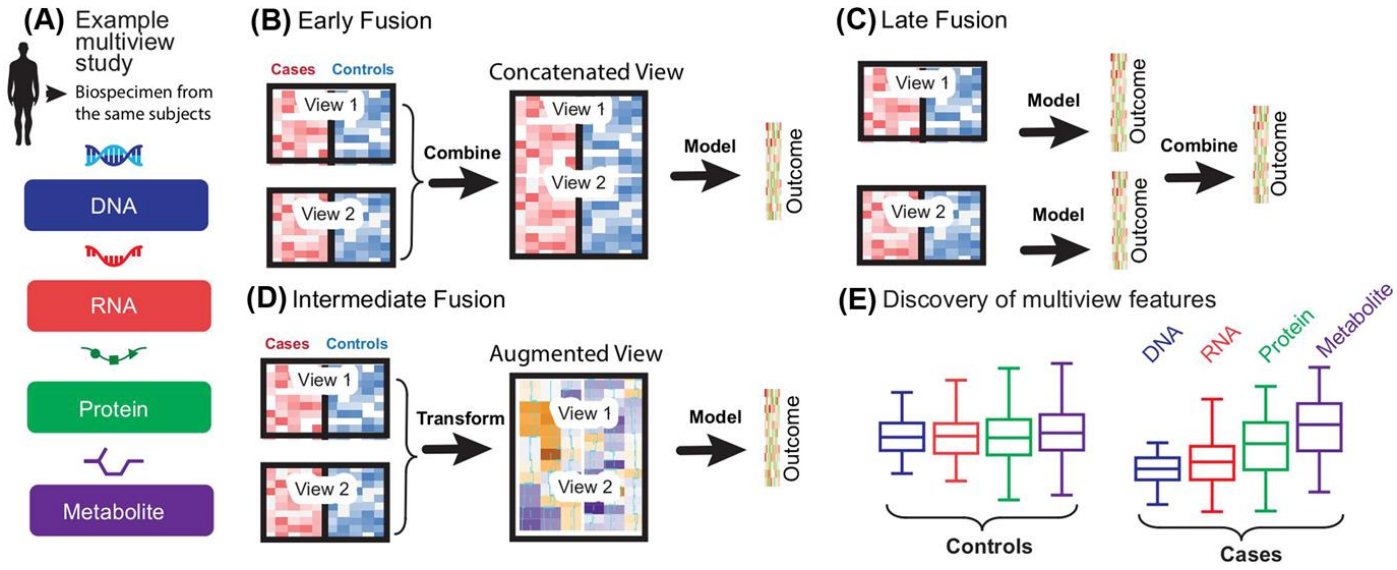
Description of Figure 12.1



In the vertical integration example, the same subjects labeled as shared anchor are measured in multiple omics. The diagram shows cases and controls for Omics 1, Omics 2 and Omics 3, with features listed from 1 to 9. In the diagonal integration example, the same subjects but labeled as No shared anchor are measured in multiple omics across non-overlapping samples. The diagram shows cases and controls for Omics 1, Omics 2 and Omics 3, arranged diagonally from top left to bottom right with features and samples indicated. In the horizontal integration example, the same multimodal features with shared anchors are measured in multiple studies. The diagram shows Omics 1 and Omics 2 separating Study 1, Study 2 and Study 3 into half, with subjects listed from 1 to 9, divided into cases and controls.

Navigate back to Figure 12.1

Description of Figure 12.2



The components in the image are: A. Example multiview study: Biospecimen from the same subjects, including D N A, R N A, Protein and Metabolite. B. Early Fusion: Cases and Controls from view 1 and 2 are combined into a Concatenated View, which is then modeled to produce an Outcome. C. Late Fusion: View 1 and View 2 are modeled separately to produce Outcomes, which are then combined to produce a final Outcome. D. Intermediate Fusion: Cases and Controls from view 1 and 2 are transformed into an Augmented View, which is then modeled to produce an Outcome. E. Discovery of multiview features: Box plots for D N A, R N A, Protein and Metabolite are shown for Controls and Cases. For controls it is arranged linearly but for cases it is arranged by increasing order.

Navigate back to Figure 12.2

Glossary

Brief Glossary of Terms Used in Metagenomics

These are concise definitions to help you understand some of the concepts at play. Resources are available to deepen your understanding from environmental microbiology, medical microbiology, experimental design, taxonomy, study design, statistics, ecology, bioinformatics, etc.

Adduct

Ions that are formed by the combination of a precursor molecule with one or more additional atoms or molecules, leading to a change in the mass-to-charge ratio observed in mass spectrometry. These adducts are important for identifying and characterizing molecules, as they aid in determining the molecular weight and structural information of the analyte.

Alpha diversity

Expression of microbial diversity at one site taking into account the number of species and their abundance. A community will have a high alpha diversity when there is a high number of species in similar relative abundance.

Aliasing

Phenomenon that renders time-varying signals indistinguishable when they are sampled below the oscillation frequencies of the signals.

Artificial intelligence (AI)

Computing technologies that are able to perform tasks that normally require human intelligence. It most often requires a learning dataset with known parameters; however, some methods enable learning differences within a dataset.

ASV

Amplicon sequencing variant (ASV) is a single DNA sequence that is used to analyze microbial communities in a microbiome. It is an alternative to operational taxonomic units (OTUs) for analyzing microbial communities. It enables to distinguish between closely related species.

Attribute

See the Microbiome attribute definition below

Autocorrelation plot

Method of analyzing time-varying dependencies, in which the correlation of a time-series with successively lagged versions of itself is plotted.

Automated experimental design

Statistical technique in which an algorithm is used to generate a design for future experiments, such as the time-points at which to sample subjects in a longitudinal study, based on certain information and objectives provided by the user.

Autoregressive model

Statistical model in which the present value of a variable is expressed as a function of its past values.

Barcode Loci

Specific regions of DNA used as markers to identify organisms.

Bayesian model

Type of statistical model, in which uncertainty in knowledge about a system takes into account both observed data as well as prior information or beliefs.

Beta diversity

Expression of the difference between sites. It takes into account the nestedness (maintenance of a subset of the population or the whole), and the turnover (species replacement) while being influenced by species richness. Page 272

Binning

Process of clustering sequences based on their nucleotide composition. This process can be enhanced using reference sequences.

Biofilm

Assemblage of microorganisms associated to themselves and/or to a substrate. They are commonly embedded within a self-produced matrix of extracellular polymeric substances.

Chimera

Artificial DNA sequence created during enzymatic amplification and/or data processing.

Collector's curve

A plot of the cumulative number of species recorded as a function of the sampling effort and/or sequencing depth. It can then be extrapolated to estimate the species richness.

Contig

Contiguous stretch of a DNA sequence composed of a set of overlapping DNA segments (reads).

Clustering algorithm

Computational method for separating data into cohesive groups according to some implicitly or explicitly specified metric.

Deep sequencing

Sequencing that provides higher coverage of the genome or transcriptome, ensuring more accurate detection of rare genetic variations or organisms.

Diploidy

Existence of two complete genome copies in one cell.

Dirichlet process

Type of stochastic process often used as a prior probability distribution in infinite mixture models.

Ecological statistics

Set of statistical tools developed and validated for macroecology that can be applied to microbiome ecology.

Feature

A variable in a dataset such as microbial taxa, protein, or metabolite. For microbial taxa, each hierarchical level can be considered a feature. The more features a dataset includes, the higher its dimensionality. Often used interchangeably with microbiome attribute.

Gene conversion

Genetic phenomenon characterized by the nonreciprocal genetic exchange between homologous sequences.

Genome resequencing

Sequencing the genome of an organism that has been previously sequenced to identify mutations, variations, or other changes.

Hidden Markov model

Type of state-space model that assumes a discrete set of states with transitions between states occurring at discrete time-steps; states are assumed to be latent or hidden variables that are inferred from data.

Host-associated

Refers to organisms or genetic material that are associated with a host organism, such as a plant, animal, or human.

Illumina sequencing

Illumina Inc. is a provider of next-generation sequencing platforms based on bridge amplification and sequence-by-synthesis.

Indel

A type of genetic variant signifying a difference in length. The term “indel“ is short for insertion-deletion, referring to how the variant appears in a sequence alignment. It is generalized as such to avoid a claim of the ancestral state as would be implied by either of the terms “insertion” or “deletion.”

Infinite mixture model

Bayesian nonparametric clustering technique, in which observations are modeled as arising from a mixture with an infinite number of clusters or components.

Information theory

Mathematical theory quantifying the amount of information in data, often involving the measure of entropy, which quantifies the degree of uncertainty involved in predicting the value of a random variable or outcome of an experiment. Page 273

In silico

Term describing scientific experiments, research, or discoveries conducted with computer modeling or computer simulation.

k-mer

A k-mer is a DNA (or amino acid) substring of length k. The occurrence of all possible substrings can be used as a k-mer frequency analysis to characterize and empirically compare genomic sequences. The optimal length k typically varies between 3 and 100 nucleotides based on the application.

Long-read sequencing

Sequencing that produces longer reads (usually over 300 base pairs), which helps with repetitive regions, genomic organization, and complex genomic structures.

Lowest common ancestor

Taxonomic or phylogenetic strategy that identifies the root of the smallest subtree of the taxonomy or phylogeny that includes all the genomes of interest.

Machine learning

Branch of artificial intelligence aimed at prediction based on patterns learned from a given input dataset, also called training dataset.

Mate pairs

Read sequences known to be in 3′ and 5′ ends of a contig. The reads may or may not overlap.

Merooligoploidy

Unequal genome dosage during cell division. When the cell division time becomes shorter than the time to replicate and segregate the chromosome, bacteria start a new round of DNA replication before the previous round has been terminated; thus the gene dosage of regions near the replication origin becomes higher than that of regions near the terminus.

Metadata

Collected information regarding the experimental parameters, as well as additional information enhancing the study at hand.

Metabolomics

Study of small molecules from microbial community, tissue, or complex sample (metabolome).

Metagenomics

Study of the genetic information coded by the DNA from microbial community, tissue, or complex sample (metagenome).

Metaproteomics

Study of the proteins of a microbial community, tissue, or complex sample (proteome).

Metatranscriptomics

Study of the genetic information coded by the RNA from microbial community, tissue, or complex sample (transcriptome).

MGI

Provider of next-generation sequencing platforms utilizing the core technology DNBSEQ based on DNA rolling nanoball (DNB) amplification and sequence-by-synthesis.

Microbiome

The community composed of microorganisms in a defined habitat. Those organisms encompass bacteria, archaea, lower eukaryotes, phages, and viruses.

Microbiome attributes

Variables that are outputted from metagenomic bioinformatic pipelines in relative abundance form, such as the relative abundance of different

microbial taxa and microbial genes.

Monoploidy

The bacteria have a single genome copy per cell, located on a single or multiple genetic elements.

Nonparametric model

Type of statistical model in which the structure of the model is not fixed, and that typically grows in size to match the complexity of data. Page 274

Oligoploidy

The bacteria have between 2 and 10 genome copies in one cell, located on a single or multiple genetic elements.

Open science

It is the practice of making scientific research, data, methods, and publications freely accessible and transparent to enable reuse by anyone.

OTU

OTU is usually defined by 16S rDNA similarities. The similarity percent chosen is often 97% for short reads and 98.5% for full-length reads. It is used in place of taxonomic organization such as species, genus, and others. This information can later be overlaid.

Paired ends

Sequence at both ends (5′ and 3′) of a segment of DNA during a sequencing process that reads from both extremities of the same DNA fragment.

Partial genome sequencing

Sequencing of only a portion of an organism's genome, used for focal regions of interest or to lower costs for large analyses.

Phylotype

Observed similarity that classifies a group of organisms together. It is rank neutral, thus it can be used for a family, genus, species, or strain based on similarities.

Polyploidy

The bacteria have more than 10 genome copies per cell, located on a single or multiple genetic elements.

PNA (peptide nucleic acids)

Synthetic molecules that bind to specific nucleotide sequences, often used to block the amplification of unwanted host DNA during PCR.

Q20 or Q30

See quality score.

Quality score, Q score, or Phred quality score

It is a parameter used to inform about the base calling accuracy. It is a property that is logarithmically related to the base calling error probability. A Q score of 20 (Q20) to a base is equivalent to the probability of an incorrect base call of 1 in 100 times (or 99% base accuracy). While a Q30 is the probability of an incorrect base call of 1 in 1000 (or 99.9% base accuracy).The higher the score is, the greater the accuracy.

Rarefaction curve

It allows the estimation of the number of species in case of a reduced sampling effort, to facilitate the comparison between communities with unequal sampling efforts.

Regression model

Statistical model in which the value of a dependent variable is expressed as a function of a set of independent variables (covariates).

Sanger sequencing

A traditional DNA sequencing method that determines the sequence of DNA by incorporating chain-terminating nucleotides during DNA synthesis.

Scaffold

When paired reads are present in two different contigs, the two contigs can be linked to form a scaffold, which is a larger noncontiguous DNA sequence. The gap size between the two contigs is estimated based on the insert size of the read pairs.

Sequence assembly

Process of putting the DNA fragments (sequence reads) back together based on sequence similarity between reads to reconstruct the original sequence.

Sequencing coverage

Mean number of times a nucleotide is sequenced in a genome (sequencing depth). Greater coverage will minimize errors due to sequencing or assembly. Page 275

Shallow sequencing

A sequencing approach that provides lower coverage of the (meta)genome or (meta)transcriptome, typically used for preliminary data or to lower costs.

Short-read sequencing

Sequencing that generates shorter reads (usually less than 300 base pairs). Includes Illumina's sequencing by synthesis approach.

Spline model

Statistical model used for regression analyses, which is specified piecewise as a series of polynomial functions that are constrained to join

together smoothly.

State-space model

Statistical model that assumes the outputs or measurements of a system depend on its state, which can change over time.

Taxonomic profiling

Identification of the organisms present in a microbial community and their relative abundances.

Whole-metagenome shotgun sequencing

In the same way as whole-genome shotgun sequencing refers to short-read sequencing of the complete genome of a single organism, whole-metagenome shotgun sequencing refers to short-read sequencing of the entire mixture of all organisms in a microbiome sample (whole-community sequencing).

Xenobiotic

Foreign chemical substance encountered by an organism. This can include pharmaceutical drugs, dietary compounds, or pollutants such as dioxins. Now it is time to have your feet dangling at the edge of leading science!

Index

Note: Page numbers followed by "*f*" and "*t*" refer to figures and tables, respectively.

A

AIC, *See* **Akaike's information criterion (AIC)**

Acanthocephalans, **84–85**

Acidibacillus, **258–259**

Ad hoc algorithm, **105**

Advanced sequencing techniques, **191**

Aerobic respiration, endosymbionts with potential for, **75–76**

Agarose gel electrophoresis, **113–114**

AI, *See* **Artificial intelligence (AI)**

Air sampling, **156–157**

Airborne microorganisms, **156–157**

Akaike's information criterion (AIC), **213**

Alistipes, **218–219**

Alphaproteobacteria, **74–75**

Alternative hypothesis, **95**

American Type Culture Collection (ATCC), **16**

Amino acids, **225–226**, **245–246**

Amphimedon queenslandica, **133**

Amplicon sequencing-based approaches, **37**

Amplicon-based methods, **55–56**

Amplified RNA (aRNA), **132**

AMR, *See* **Antimicrobial resistance (AMR)**

Anaerobic protists, **75**

Analysis of covariance (ANCOVA), **100**

Analysis pipelines, **39**

Analytical chemistry techniques, **225**

Analytical peaks, **226–227**

Analytical platforms, **232**

Anchors, **250–251**

ANCOVA, *See* **Analysis of covariance (ANCOVA)**

Ancylostoma duodenale, **81–82**

ANI, *See* **Average nucleotide identity (ANI)**

Animal research, **6**

Anoplocephala perfoliata, **81–82**

Antibiotics, **209–210**

Antimicrobial resistance (AMR), **47**, **147**, **172–173**

Arctic environments, **55**

Arithmetic precision, **213–214**

aRNA, *See* **Amplified RNA (aRNA)**

Artificial intelligence (AI), **39–40**, **60–61**, **192–193**, **245**

Ascaris, **84**

Ascaris lumbricoides, *See* **Roundworms (*Ascaris lumbricoides*)**

ASD, *See* **Autism spectrum disorder (ASD)**

Assembled metagenomic data, **193–194**

Assembly-based approaches, **253**

Associated analytic methodologies, **27**

Associated data, **4**

ATCC, *See* **American Type Culture Collection (ATCC)**

Attribute-specific model, **213–214**, **213**

Autism spectrum disorder (ASD), **259**

Automated instruments, **56**

Automated ribosomal intergenic spacer analysis, **32–33**

Average nucleotide identity (ANI), **198**

B

Bacillus

B. anthracis, **48**, **52–53**

B. subtilis, **16**

B. thuringiensis, **52–53**

BactDrop, **134**

Bacteria, **76–77**, **77–78**

Bacteriology, **16**

Bacteroidetes, **74–75**

Barcode loci, **30–32**

Barroeca monosierra, **77–78**

BAs, *See* **Bile acids (BAs)**

Batch effect, **10**

Bayesian information criterion (BIC), **212**

Betaproteobacteria, **74–75**

BGCs, *See* **Biosynthetic gene clusters (BGCs)**

BIC, *See* **Bayesian information criterion (BIC)**

Bicosta minor, **77–78**

Bifidobacterium, **49**

Bile acids (BAs), **243–244**

Binning

- contig length on, ***193f***
- definitive method for, **191–192**
- methods, **175–176**, **191–192**
- probabilistic nature of, **195**

Biogeography, concept of, **12–13**

Page 278 Bioinformatic pipelines, **58**

Bioinformatics tools, **152**, **152–153**

Biological pipeline, **2–3**

Biological systems, **39**

Biomarkers, **106**

Biosafety Level 3 (BSL-3), **50–51**

Biostatisticians, **107**

Biosynthetic gene clusters (BGCs), **172–173**

Bioterrorism, **51**

Biothreat, **47**

agents, **56**

Bligh method, **231**

Blocked rRNA, enrichment of mRNA by, **126**

Botulinum neurotoxins, **49**

Breast cancer, **258–259**

Brown stomach worm, **81–82**

Brucella, **48**, **50**

B. anthracis, **49**

B. melitensis, **49**

Brugia malayi, **83**

BSL-3, *See* **Biosafety Level 3 (BSL-3)**

Burholderia cepacia, **58**

Burkholderia sp., **129–130**, **131–132**

B. pseudomallei, **58**

BurkHostGEN, **60–61**

C

Candidatus Azoamicus ciliaticola, **75–76**

Carbohydrate-active enzyme database (CAZy), **14**

Carbohydrates, **225–226**

Cardiovascular disease, **225**

Cas proteins, *See* **CRISPR-associated proteins (Cas proteins)**

cDNA, *See* **Copy DNA (cDNA)**

Cellular metabolic processes, **230**

CheckM, **195**

CheckM1, **195**

CheckM2, **196**

Cherry-picking, **99**

Chi-square test, **100**, **105–106**

Chimera formation, **30**

Chipmunk (*Tamias amoenus cratericus*), **84**

Chlamydiae, **74–75**

Choanoflagellates microbiome and origin of metazoans, **77–78**

CHOW, *See* **Commercially available monkey feed (CHOW)**

Chromatographic RT, **226–227**

CIOMS, *See* **Council for International Organizations of Medical Sciences (CIOMS)**

Ciona, **79–80**

Circularized MAGs (cMAGs), **194**

Clade designation, **57–58**

Climate change, **55**

Clinical research, **143**

Clostridia, **102–103**

Clostridium, **258–259**

C. botulinum, **49**, **53–55**

Clustered regularly interspersed short palindromic repeats (CRISPR), **124–125**

CRISPR/Cas9-based rRNA depletion, **124–126**

cMAGs, *See* **Circularized MAGs (cMAGs)**

Cohort description, **209–210**

Colorectal cancer (CRC), **258–259**

Commensalism, **67–68**

Commercially available monkey feed (CHOW), **98–99**

Community DNA, **119–120**

Community-specific rRNA depletion method, **119–120**

Complete metagenome-assembled genomes, **146–148**

Complex ecosystem, **243**

Compositional data, **19**

Comprehensive variant detection from single technology, **146–148**

Computational efforts, **39–40**

Computational techniques, advanced, **120**

Computational tools, **191**

Computing resources, **14**

Conexibacter woesei, **244**

Contextual analysis, **13**

Contig, **14**, **172–173**
length, **194**
Contrast mixotrophs, **74**
Cooperative learning, **255–256**
Copy DNA (cDNA), **145**
library
clone libraries, **109**
normalization, **130–131**
selective priming in, **128–130**
synthesis, **110**
Corynebacterium pseudobuberculosis, **114**
Council for International Organizations of Medical Sciences (CIOMS), **6**
Coxiella burnetii, **48–49**
Cramer's Phi statistic, **103–104**
CRC, *See* **Colorectal cancer (CRC)**
CRISPR, *See* **Clustered regularly interspersed short palindromic repeats (CRISPR)**
CRISPR-associated proteins (Cas proteins), **124–125**
Cross-sectional study design, **5**
Cultivation-independent analysis, **1**
Page 279 Culture-based techniques, **5**, **47**
Culture-independent molecular methods, **67**
Cutting-edge platforms, **60**
Cyanobacteria, **74–75**

Cyberbiosecurity, **60–61**

D

DASH, *See* **Depletion of abundant sequences by hybridization (DASH)**

Data

- acquisition, **232–233**
- analysis, **13–14**
- availability, **60**
- dredging, **99**
- hypothesis-driven research, **226–227**
- processing and annotation, **233–234**
- publication and dissemination, **17–18**

DBS, *See* **Dried blood spot (DBS)**

Decision-making process, **11**

Deep neural networks, **40**

Denaturing gradient gel electrophoresis (DGGE), **32–33**

Denitrification respiration, endosymbionts with potential for, **75–76**

Dental pulp, **53–55**

Depletion of abundant sequences by hybridization (DASH), **124–125**

Desert Locust (*Schistocerca gregaria*), **7–8**

Design of experiments (DOEs), **120–121**

DGGE, *See* **Denaturing gradient gel electrophoresis (DGGE)**

Diagonal integration, **252**

Diatom, host genotype structures microbiome of, **76–77**

Diet-microbiome associations, application of IDSF to assess, **218–219**

Dirichlet-multinomial model (DM model), **103**

hypothesis test, **103–104**

Discussed above, **102**

Distribution components, **216**

Distribution-specific models, **212–213**, **212**, **213**

DIY, *See* **Do-it yourself (DIY)**

DM model, *See* **Dirichlet-multinomial model (DM model)**

DNA

approaches, **145–146**

Data Bank of Japan, **17**

DNA-seq, **220**

technologies, **143**

extraction protocols, **194–195**

viruses, **124–125**

Do-it yourself (DIY), **119**, **120**

DOEs, *See* **Design of experiments (DOEs)**

Double stranded cDNA, **130**

Downsampling experiments, **180–181**

Downstream binning steps, **173–174**

Dried blood spot (DBS), **230**

Drug safety, **49–50**

DSN, *See* **Duplex-specific nuclease (DSN)**

Dual RNA-seq, ribosomal RNA depletion in, **132–133**

Dual use potential, **60–61**

Duplex-specific nuclease (DSN), **130**

Dyer method, **231**

Dynamic ecosystem, **243**

E

Earth, **7**

eukaryogenesis, **71**

Earth Microbiome Project, **39**

Echinoderms, **79**

EDTA, *See* **Ethylenediaminetetraacetic acid (EDTA)**

EMBR-seq, *See* **Enrichment of mRNA by blocked rRNA-seq (EMBR-seq)**

Emerging infectious agents, **47**

Emerging threats, **55**

Endosymbiont–microbiome continuum, **70–71**

Endosymbionts with potential for denitrification and aerobic respiration, **75–76**

Endosymbioses, **72–73**

in origin of eukaryotes and mitochondria, **71–72**

Endosymbiotic relationships, **68**

Enrichment of mRNA by blocked rRNA-seq (EMBR-seq), **115–116**, **126–127**

Enterobacter cloacae, **218–219**

Environmental health, applications of metagenomic nanopore sequencing for, **154–159**

Environmental microbiomes, exploring depths of diversity in, **156–159**

Environmental nucleic acids, **7**

Environmental samples, **4**, **7**, **8**

Environmental transcriptomics, **109**

EPI2ME platform, **153**

Epigenetic data, **148**

Epistasis, **102**

Escherichia coli, **15–16**, **113–114**

Ethylenediaminetetraacetic acid (EDTA), **228**

Eubacterium ramulus, **218–219**

Eukaryotes, endosymbioses in origin of, **71–72**

Eukaryotic mRNAs, **109–110**

Eukaryotic phylogenetic diversity, **78**

European Nucleotide Archive (Europe), **17**

Page 280 Exposure-attribute model, **215**

Exposure-attribute statistical model, **214–215**

Extraction method, **231–232**

F

Faecalibacterium, **259**

Falling sequencing costs, **36–37**

Fasciola hepatica, **81–82**

Fatty acids, **225–226**

FECA, *See* **First eukaryotic common ancestor (FECA)**, **Fusion event common ancestor (FECA)**

Fecal metabolomes, **228–229**

Fecal samples, metagenomic description of, **210**

Fingerprinting, **34–36**

Finite mixture modeling, **208–209**, **216–217**

Firmicutes, **16**

First eukaryotic common ancestor (FECA), **72**

FL, *See* **Full-length (FL)**

Food and Drug Administration guidelines, **106**

Food safety, **49–50**

Fragmentation, **193–194**

Francisella, **49**, **51–52**

F. tularensis, **50**

F-like organisms, **51–52**

Full-length (FL), **165**

gene amplicon microbiome sequencing, **165–168**

16S rRNA amplicon sequencing, ***166f***

Kinnex 16S rRNA kit results, ***168f***

Fusion event common ancestor (FECA), **72**

Fusobacterium nucleatum, **258–259**

G

Gammaproteobacteria, **74–75**

Gas chromatography-mass spectrometry (GC-MS), **232**

Gastric cancer, **258–259**

Gastrointestinal tracts (GI tracts), **67**, **81–82**, **207**

Gaussian distributions, **219**

GC-MS, *See* **Gas chromatography-mass spectrometry (GC-MS)**

Gel electrophoresis, size selection by, **113**

Gene

- conversions, **71**
- structure, **109**

Genome centric metagenomics

- key differences between MAGs and isolate genome, **199–200**
- MAGs, **191–192**
  - construct, **192–195**
  - quality standards, **195–197**
- naming and organizing uncultivated, **197–199**

Genome taxonomy database (GTDB), **198**

Genome Taxonomy Database Toolkit (GTDB-Tk), **175–176**, **199**

Genome-centric metagenomics, **191**

Genome(s), **173–174**

copies, **15–16**

size, **15**

structure, **193–194**

GeoChip, **33**

German Collection of Microorganisms and Cell Cultures (DSMZ), **16**

GI tracts, *See* **Gastrointestinal tracts (GI tracts)**

Giganucleotides (Gnt), **210**

Global Parasite Project, **82**

Gnt, *See* **Giganucleotides (Gnt)**

Gradient boosted decision trees, **196**

"Green Belt" area, **156–157**

GTDB, *See* **Genome taxonomy database (GTDB)**

GTDB-Tk, *See* **Genome Taxonomy Database Toolkit (GTDB-Tk)**

Gut microbiome studies, **67**

H

Haemonchus contortus, **81–82**, **84–85**

Haloquadratum walsbyi, **244**

Halorhabdus sp., **114–115**

Health Professionals Follow-up Study (HPFS), **209–210**

"Healthy" soil communities, **29**

Helicobacteraceae, **258–259**

Helminth-infected humans, **82–83**

Helminths, **81–85**

Heteroxynema cucullatum, **84**

Hidden Markov models (HMMs), **195**

HiFi full-length amplicon sequencing, **166–167**

HiFi reads, **164**

HiFi-MAG-Pipeline, **175–176**

HiFi-specific assembly methods, **176–177**

High molecular weight DNA, **194–195**

High-dimensional datasets, **30**

High-performance liquid chromatography, **114**

High-purity, **194–195**

High-quality

genomes, **163–164**

step closer high-quality datasets to, **10–11**

High-quality-MAGs (HQ-MAGs), **176–177**

High-throughput sequencing technologies, **34**

Page 281 Higher sequencing costs, **36–37**

Higher throughput assessment technologies, **207**

HILIC, *See* **Hydrophilic interaction liquid chromatography (HILIC)**

HIPAA Privacy Rule, **6**

HiSeq paired-end sequencing platform, **210**

HMMs, *See* **Hidden Markov models (HMMs)**

HMP, *See* **Human Microbiome Project (HMP)**

HMR, *See* **Human/Mouse/Rat (HMR)**

Holobiont, **68–69**

HOMD, *See* **Human Oral Microbiome (HOMD)**

Homemade methods, **120–121**

Homoscedasticity, **100**

Hookworms, **81–82**

Horizontal integration, **250–251**, **252**

Host genotype structures microbiome of diatom, **76–77**

Host nucleic acids, **7**

Host omics, **250**

Host population studies, **220–221**

Host-associated microbiomes, **4**, **37**

Host–microbiome interactions, **68–69**

Host–microbiota associations, **70–71**

Host–microbiota interactions, **81**

Hot Start Taq DNA, **30**

HPFS, *See* **Health Professionals Follow-up Study (HPFS)**

HQ-MAGs, *See* **High-quality-MAGs (HQ-MAGs)**

Human exposures, **207**

Human gastrointestinal microbiome, **207**

Human gut microbiome, **56**

Human gut sample, **179–180**

Human health

applications of metagenomic nanopore sequencing for, **154–159**
microbiome multiomics, **244**
Human microbiome, **163**
Human Microbiome Project (HMP), **7**, **100**, **255–256**
consortium, **7**
R HMP package, **105**
Human Oral Microbiome (HOMD), **14**
Human sequences, **210**
Human stool microbiome, **220–221**
Human/Mouse/Rat (HMR), **133**
Hybrid approaches, **227**
Hydrophilic interaction liquid chromatography (HILIC), **232**
Hypothesis testing, **95**, **105–106**
hypothetical normal distributions of two groups, ***96f***
microbiome data, **97**

I

IBD, *See* **Inflammatory bowel disease (IBD)**
ICB, *See* **Immune checkpoint blockade (ICB)**
ICNP, *See* **International Code of Nomenclature of Prokaryotes (ICNP)**
IDSF, *See* **Intelligent decision support framework (IDSF)**
iHMP, *See* **Integrative Human Microbiome Project (iHMP)**
Illumina HiSeq paired-end sequencing platform, **210**

Illumina's NovaSeq, **36–37**

Immune checkpoint blockade (ICB), **258–259**

In situ hybridization, **134**, **134–135**

In situ RNase H digestion, **134**

In situ rRNA depletion, **135–136**

In vitro transcription (IVT), **126–127**, **135–136**

In-house methods, **120–121**

Independent endosymbiosis, **73**

Inflammatory bowel disease (IBD), **243–244**

Information-losing transformations, **103**

Institutional Review Boards (IRBs), **6**

Integrated multiomics

- approaches, **39**
- datasets, **255–256**

IntegratedLearner, **255–256**

Integrative Human Microbiome Project (iHMP), **243–244**, **250**

Intelligent decision, **215**

Intelligent decision support framework (IDSF), **209**

- application of IDSF to assess diet-microbiome associations, **218–219**
- component 1 of, **211–213**
  - attribute-specific model identification, **213**
  - distribution-specific models identification, **212–213**
  - methodologic rationale, **211–212**

component 4 of, **214–215**

components 2 and 3 of, **213–214**

arithmetic precision, **213–214**

optimization, **214**

components comprising, ***211f***

in context, **219–221**

methodology, **210–211**

observed distributions of microbiome attributes as identified using, **217–218**

rationale for, **216–217**

Page 282 "Intelligent" component probability, **216**

Internal transcribed spacer regions (ITS regions), **165–166**

International Code of Nomenclature of Prokaryotes (ICNP), **197**

International Nucleotide Sequence Database Collaboration, **197**

Intraspecies microbial diversity, **16**

Ion chromatography, **232**

IRBs, *See* **Institutional Review Boards (IRBs)**

Isolate genome, differences between MAGs and, **199–200**

ITS regions, *See* **Internal transcribed spacer regions (ITS regions)**

IVT, *See* ***In vitro* transcription (IVT)**

J

JIVE, *See* **Joint And Individual Variation Explained (JIVE)**

Joint And Individual Variation Explained (JIVE), **255–256**

K

Kinnex 16S rRNA kit, **166–167**, **168*f***

L

Lachnoclostridium, **258–259**

Lachnospiraceae, **258–259**

Last eukaryotic common ancestor (LECA), **72**

LC-MS, *See* **Liquid chromatography-mass spectrometry (LC-MS)**

LECA, *See* **Last eukaryotic common ancestor (LECA)**

Legacy sequencing technologies, **149**

Lipids, **225–226**

Liquid chromatography, **232**

Liquid chromatography-mass spectrometry (LC-MS), **232**

Lokiarchaeota, **71–72**

Long-read PacBio HiFi sequencing

- full-length gene amplicon microbiome sequencing, **165–168**
- metagenome assembly, **173–182**
- taxonomic and functional profiling of metagenomic reads, **168–173**

Long-read sequencing, **34**, **164**, **174–175**, **194**

Longitudinal analysis, **38**

Longitudinal data, **34–36**

Low-quality MAGs, **196–197**

Lower respiratory infections (LRIs), **154–155**

LRIs, *See* **Lower respiratory infections (LRIs)**

M

M3-seq, *See* **Massively parallel, multiplexed, microbial sequencing (M3-seq)**

Machine learning, **234–236**, **245**

MAGs, *See* **Metagenome-assembled genomes (MAGs)**

Marine deposits, **244**

Marker-based strategy, **201**

MAS-Seq, *See* **Multiplexed array sequencing (MAS-Seq)**

Mass spectrometry (MS), **225**, **248–249**

Massively parallel, multiplexed, microbial sequencing (M3-seq), **135–136**

MATQ-seq, **135**

Mean of distribution, **103–104**

Medium-quality MAGs, **196–197**

MEGAN-LR, **171**

Men's lifestyle validation study (MLVS), **209–210**

- cohort description, **209–210**
- metagenomic description of fecal samples, **210**

Messenger RNA (mRNA), **109**

Meta-analyses, **250–251**

Metabolite(s), **225**

- annotation, **233–234**

detection, **232–233**

extraction, **231**

Metabolome, **225–226**

Metabolomics, **225–226**, **243**, **245–248**, **245**, **247*t***

data processing, **233**

and annotation, **233–234**

DBS, **230**

instruments used, **232–233**

integration of, **234–236**

sample

extraction, **231–232**

handling considerations for, **227–229**

society, **225–226**

standards initiative, **233–234**

targeted *vs*. untargeted, **226–227**

tissue, **230**

Metadata, **4**, **12–13**

Metagenome assembly, **173–182**, **174*f***, **177*f***

Metagenome-assembled genomes (MAGs), **14**, **75–76**, **146–147**, **173–174**, **177–179**, **178*f***, **180*f***, **191**, **191–192**, **192*f***

construction, **192–195**, **193*f***

MAGs and isolate genome, differences between, **199–200**, **200*f***

quality standards, **195–197**

Metagenomic data, **97–98**

Page 283 analysis, **152–153**

compare all taxa across groups simultaneously, **102–105**

compare diversity across groups, **98–99**

power curves for testing mean Shannon diversities, ***100f***

Shannon diversity index, ***99f***

comparing single taxon across groups, **99–102**

contingency table showing counts of samples, ***101t***

power curves for present/absent contingency, ***102f***

future research, **105–107**

power curve for two-sided *t*-test, ***97f***

standard format of microbiome sequence read data by sample and taxa, ***98t***

Metagenomic nanopore sequencing

applications for human and environmental health, **154–159**

exploring depths of diversity in environmental microbiomes, **156–159**

transformative potential of nanopore sequencing in respiratory metagenomics, **154–156**

metagenomic data analysis, **152–153**

nanopore sequencing, **152**

sample extraction and library preparation, **150–152**

DNA library preparation workflow, ***151f***

workflow, **150–153**

Metagenomic next generation sequencing (mNGS), **47**, **48**, **50**, **53**, **55–56**

Metagenomic(s), **1–3**, **14**, **40**, **67**, **143**, **163**, **243**

availability of commercial kits/instruments, **56**

challenges, **59–60**

choanoflagellates microbiome and origin of metazoans, **77–78**

climate change and emerging threats, **55**

SNP-based heatmap, ***54f***

clinical utility, **47–49**

contamination by host and/or environment, **7–8**

data

analysis, **13–14**

publication and dissemination, **17–18**

dataset, **5**

description of fecal samples, **210**

details frequently overlooked, **14–16**

dual use potential and cyberbiosecurity, **60–61**

endosymbionts with potential for denitrification and aerobic respiration, **75–76**

endosymbioses

in origin of eukaryotes and mitochondria, **71–72**

and plastids evolution, **72–73**

epidemiological investigation, **53–55**

equation between hypothesis, sequencing depth, and financial matters, **11–12**

ethical and legal issues, **6–7**

examining status quo, **19**

fieldwork, **50–51**

food and drug safety, **49–50**

helminths, **81–85**

host genotype structures microbiome of diatom, **76–77**

integrating scales, **18**

integration of, **234–236**

laboratory developed workflows, **56**

metadata, **12–13**

microbiomes

    and symbioses across invertebrates, **79–81**

    within microbiomes, **81–85**

mNGS approaches, **55–56**

multiomics approaches in, **245–250**

nanopore sequencing enhance metagenomic research, **146–150**

sample

    acquisition and storage, **8–10**

    extraction, **10–11**

    processing, **56–57**

sequence analysis, **57–59**

    mNGS datasets may contain pathogen genomes, ***58f***

sequencing, **11–12**, **153–154**, **156**
approaches, **67**
workflow, **150–151**
statistical analysis plan, **4–5**
study design, ***3f***, **4**
surveillance, **51–53**
symbioses in protists, **73–75**
taxonomic and functional profiling of, **168–173**, ***172f***
workflows, **56**
Metaorganism, **68–69**
Metaproteomics, **243**, **245**, **248–249**, ***249t***
Metatranscriptomics, **27**, **109**, **243**, **245**, **248**
metaWGS, **175–176**
Metazoans, **79**
choanoflagellates microbiome and origin of, **77–78**
Page 284 *Methanocaldococcus jannaschii*, **128–129**
Methanogen endosymbionts, **75**
Methylation, **145–146**
5-methylcytosine (5mC), **153**
Microarray technology, **33**
MICROBExpress, **118**, **131–132**
bacterial mRNA enrichment kit, **116**
Microbial communities, **32**, **32–33**, **67**, **245–246**, **250**

Microbial consensus systematic, **15**

Microbial dark matter, **196**

Microbial dependence continuum, **69–70**

Microbial diversity, **163–164**, **197**

Microbial genomes, **148**

Microbial metagenomics, **27**

Microbial research, **143**

Microbial transcriptomics, **109**, **109**

Microbial-omics analysis, **28–29**

Microbiome attributes, **207**, **208**, **213**, **219**

- excellent resource for studying distributional profiles of, **209–210**
- observed distributions of microbiome attributes as identified using IDSF, **217–218**

Microbiome characterization methods

- expectations and predictions, **38–40**
- shifting methodologies and large datasets, ***31f***, **34–38**
- traditional technologies scaled up and new-omics approaches, **29–34**

Microbiome(s), **13**, **29–30**, **29**, **38**, **69–70**, **81**, **105**, **243**

- choanoflagellates microbiome and origin of metazoans, **77–78**
- data, **102–103**, **211–212**, **219**, **220**
- ecosystem, **208**
- endosymbionts with potential for denitrification and aerobic respiration, **75–76**

endosymbioses
in origin of eukaryotes and mitochondria, **71–72**
and plastids evolution, **72–73**
helminths, **81–85**
host genotype structures microbiome of diatom, **76–77**
microbiomes
and symbioses across invertebrates, **79–81**
within microbiomes, **81–85**
microbiome-based hypotheses, **207–208**
multiomics
frameworks for, **250–258**
success stories in, **258–259**
relative abundance data, **219**
research, **163**
sequence datasets, **209**
studies, **37**, **74**, **208**
methodologies for determining probability distributions in, **207–209**
symbioses in protists, **73–75**
taxonomic sequence counts data, **105–106**
Microbiota transplant therapy (MTT), **259**
MIMARKS, *See* **Minimal information about marker gene sequence (MIMARKS)**

MIMS, *See* **Minimal information about metagenome sequence (MIMS)**

Minimal information about marker gene sequence (MIMARKS), **17**

Minimal information about metagenome sequence (MIMS), **17**

Minimum information about any (X) sequence (MIxS), **17**

Mitochondria, endosymbioses in origin of, **71–72**

MIxS, *See* **Minimum information about any (X) sequence (MIxS)**

Mixture probabilities, **211–212**

MLVS, *See* **Men's lifestyle validation study (MLVS)**

mNGS, *See* **Metagenomic next generation sequencing (mNGS)**

Modern genomics, **16**

Modification of transcripts (mRNA), **225–226**

- cleavage of, **117–118**
- degradation, **115**
- enrichment of mRNA by blocked rRNA, **126**
- high-GC fraction of, **130–131**

MOFA, *See* **Multiomics factor analysis (MOFA)**

Molecular biology, **207**

Molecular cloning sequencing, **32–33**

Molecular techniques, **5**

5′-monophosphates, **114–115**

Motor protein, **145**

mpox, **57–58**

mRNA, *See* **Messenger RNA (mRNA)**, **Modification of transcripts (mRNA)**

MS, *See* **Mass spectrometry (MS)**

MTT, *See* **Microbiota transplant therapy (MTT)**

Multimodal AI, **245**

Multimodal data fusion

frameworks of, **251*f***

three paradigms of, **254*f***

Multimodal tools, **244–245**

Page 285 Multinomial model, **105**

Multiomics

approaches, **244–245**, **258**

in metagenomics, **245–250**

characterization, **84–85**

datasets, **253**

frameworks for microbiome multiomics integration, **250–258**

unsupervised multiomics integration methods, **256*t***

host omics, **250**

integration, **250–251**

metabolomics, **245–248**

metaproteomics, **248–249**

metatranscriptomics, **248**

success stories in microbiome multiomics, **258–259**

supervised microbiome multiomics integration methods, **257*t***

Multiomics factor analysis (MOFA), **255–256**

Multiplexed array sequencing (MAS-Seq), **166–167**

Multiplexing, **151**

Multisite collection, **38**

Multistudy integration, **250–251**

Multiview clustering (MVC), **255–256**

Mutualism, **67–68**

MVC, *See* **Multiview clustering (MVC)**

Mycobacterium smegmatis, **119**

N

N4-methylcytosine (4mC), **153**

N6-methyladenine (6 mA), **153**

Nanopore sequencing, **145**, **148**, **152**, **153**

enhancing metagenomic research, **146–150**

complete metagenome-assembled genomes from single technology, **146–148**

comprehensive variant detection from single technology, **146–148**

DNA structure description automatically generated with medium confidence, ***147f***

on-demand metagenomic sequencing, **149–150**

portable Oxford Nanopore MinION sequencers, ***150f***

real-time sequencing and rapid workflows, **149**

kits, **151**

nanopores, **144–145**

technology, **149**

transformative potential of nanopore sequencing in respiratory metagenomics, **154–156**

work, **143–146**, ***144f***

Nanopores, **146–147**, **147**

technology, **156**

National Collection of Industrial, Food and Marine Bacteria (NCIMB), **16**

National Human Genome Research Institute (NHGRI), **6**

National Institute of Health, **6**

National Library of Medicine, **13**

Natural DNAses, **9**

NCBI (USA), **17**

NCD Risk Factor Collaboration (NCD-RisC), **14–15**

NCD-RisC, *See* **NCD Risk Factor Collaboration (NCD-RisC)**

NCIMB, *See* **National Collection of Industrial, Food and Marine Bacteria (NCIMB)**

ncRNA, *See* **Regulatory noncoding RNA (ncRNA)**

Necator americanus, **81–82**

Nematoda, **84**

Neorickettsia, **83–84**

Network analysis, **234–236**

Neural networks, **196**

Neurological disorders, **225**

multiomics, **259**

New-omics approaches, **29–34**

Next-generation sequencing (NGS), **4–5**, **13**, **47**, **67**, **97**, **165**

availability of, **1**

NGS, *See* **Next-generation sequencing (NGS)**

NHGRI, *See* **National Human Genome Research Institute (NHGRI)**

Nonparametric model, **107**

Normal distribution, **95**

Normalization functions, **219**

Not-so-random primers, **129**, **129–130**

Nuclear magnetic resonance spectroscopy, **225**

Nucleic acids, **1**

Nucleic-acid-based “-omics” studies, **27–28**

Nucleotides, **225–226**

Null hypothesis, **69–70**, **95**, **98–99**

O

Observed symbiotic interactions, **68**

Old data, **19**

Oligo(dT) chromatography, **128**

Omics

data

Page 286 application of IDSF to assess diet-microbiome associations, **218–219**

component 1 of IDSF, **211–213**

component 4 of IDSF, **214–215**

components 2 and 3 of IDSF, **213–214**

IDSF in context, **219–221**

IDSF methodology, **210–211**

men's lifestyle validation study, **209–210**

methodologies for determining probability distributions in microbiome studies, **207–209**

observed distributions of microbiome attributes as identified using IDSF framework, **217–218**

rationale for IDSF, **216–217**

studies

contamination by host and/or environment, **7–8**

data analysis, **13–14**

data publication and dissemination, **17–18**

details frequently overlooked, **14–16**

equation between hypothesis, sequencing depth, and financial matters, **11–12**

ethical and legal issues, **6–7**

examining status quo, **19**

integrating scales, **18**

metadata, **12–13**

metagenomics, **1–3**
sample acquisition and storage, **8–10**
sample extraction, **10–11**
sequencing, **11–12**
statistical analysis plan, **4–5**
study design, ***3f***, **4**
On-demand metagenomic sequencing, **149–150**
On-site metagenomic nanopore sequencing, **149–150**
Onchocerca volvulus, **83**
One-component distributions, **208–209**
ONT, *See* **Oxford Nanopore Technologies (ONT)**
Open-source analytical tools, **60–61**
Operational taxonomic units (OTUs), **37**
Orchis provincialis, **2**
Organic acids, **225–226**
Organism-specific subtraction system, **119**
Orthopoxvirus, **53**
OTUs, *See* **Operational taxonomic units (OTUs)**
Ovation Prokaryotic RNA-seq System, **129–130**
Overdispersion, **103–104**, **105**
Oxford Nanopore, **149**
sequencing
platform, **52–53**

protein nanopores, **143**

Oxford Nanopore Promethion R10.4, **194**

Oxford Nanopore Technologies (ONT), **50–51**, **143**, **164**

applications of metagenomic nanopore sequencing for human and environmental health, **154–159**

metagenomic

nanopore sequencing workflow, **150–153**

sequencing or 16S sequencing, **153–154**

nanopore sequencing enhance metagenomic research, **146–150**

nanopore sequencing work, **143–146**

Oxford Nanopore MinION™ flow cell, ***145f***

P

p-hacking, **99**

P-values, **106**

PacBio HiFi sequencing, **164**

PacBio®, *See* **Pacific Biosciences® (PacBio®)**

Pacific Biosciences HiFi generate long contig-length reads, **194**

Pacific Biosciences® (PacBio®), **164**

Pan-riboPOOLs, **118–119**

Parametric models, **107**

Parametric statistical models, **105**

Parasite Microbiome Project (PMP), **82**

Parasitism, **67–68**

Parasitology, **82**

Parkinson's disease (PD), **259**

PCA, *See* **Principal component analysis (PCA)**

PCR, **37**

PD, *See* **Parkinson's disease (PD)**

Peptides, **248–249**

Permafrost-preserved microorganisms, **55**

Phenotypes, **163**

Phoronida, **80–81**

Phoronids, **80**, **80–81**

5′phosphate-dependent exonuclease digestion, **114–115**

Photosynthetic cyanobacterium, **72–73**

Phycosphere, **76–77**

PhyloChip, **33**

Pinworms, **84**

Placozoa, **80**

Plagiopylea, **75–76**

Plastids evolution, **72–73**

PMP, *See* **Parasite Microbiome Project (PMP)**

Pocket-sized MinION device, **152**

3′polyadenylation against rRNA, **128**

Page 287 Polyploidy, **15–16**

Porphyromonadaceae, **258–259**

Post hoc Cas9-based depletion, **135**

Post hoc RNase H digestion, **124**

after *in vitro* transcription, **135–136**

Post hoc test, **107**

Power calculation, **103–104**

Precision model, **214**, **215**

Prevotella, **259**

Principal component analysis (PCA), **255–256**

Probability distribution, **207–209**, **208–209**

Probability model, **105**

ProBac-seq, *See* **Probe-based bacterial RNA-seq (ProBac-seq)**

Probe-based bacterial RNA-seq (ProBac-seq), **134–135**

Probes-rRNA hybridization, **120**

Prokaryotic symbiosis, **74–75**

PromethION sequencers, **152**

Protein-coding genes, **196**

Protists, symbioses in, **73–75**

Pseudomonadota, **74–75**

Pseudomonas, **244**

P. aeruginosa, **122**

Psychrobacter, **244**

Pyrococcus horikoshii, **128–129**

Q

Quality assessment, **195**

Quality metrics, **198**

Quality score, **175–176**

R

RamDA-seq, *See* **Random displacement amplification sequencing (RamDA-seq)**

Random displacement amplification sequencing (RamDA-seq), **135**

Random variable, **98**

Rapid workflows, **149**

Rattus norvegicus, **7–8**

Raw metagenomic data, **192–193**

Real-time sequencing, **149**

Recurrent urinary tract infections (rUTIs), **259**

RED, *See* **Relative evolutionary divergence (RED)**

Reference genomes, **152–153**

Regression parameter, **211–212**

Regressor value, **211–212**

Regulatory noncoding RNA (ncRNA), **109**

Relative evolutionary divergence (RED), **198**

Removal efficiency, **116–117**

Respiratory metagenomics, transformative potential of nanopore sequencing in, **154–156**

Retention time (RT), **226–227**

Retrospective analysis, **53–55**

Reverse primers, **119–120**

Reverse-phase liquid chromatography, **232**

Rhodopseudomonas palustris, **129**

Ribo-Zero, **118**

kit, **117**

rRNA Removal Kit, **116–117**

RiboMinus, **118**

RiboMinus Transcriptome Isolation Kit, **117–118**

RiboRid, **123**

Ribosomal RNA (rRNA), **109**, **109–110**, ***111t***, **130**, **153**

5′ phosphate-dependent exonuclease digestion, **114–115**

biased 3′ polyadenylation against rRNA, **128**

cDNA library normalization, **130–131**

clustered regularly interspersed short palindromic repeats/Cas9-based rRNA depletion, **124–126**

combination of methods, **131–132**

denaturing high-performance liquid chromatography, **114**

depletion methods, ***111t***

in dual RNA-seq, **132–133**

enrichment of mRNA by blocked rRNA, **126**

RNase H digestion, **121–124**

selective priming in cDNA synthesis, **128–130**

in single-cell RNA-seq, **134–136**

size selection by gel electrophoresis, **113**

subtractive hybridization, **115–121**

Rickettsia, **51–52**

R. japonica, **53–55**

RIN, *See* **RNA integrity number (RIN)**

RMg workflow, **156**

RNA, **11**, **114**

approaches, **145–146**

chromatography, **114**

high-performance liquid chromatography for, **114**

physical separation of, **114**

poly-A tails, **11**

quantity, **110–113**

RNA-seq, **109–110**

analysis, **51–52**

data, **220**

RNAses, **9**

viruses, **156**

RNA integrity number (RIN), **118–119**

RNase H

Page 288 digestion, **121–124**

method, **123**

RNase H-based method, **122**

Roche's 454 pyrosequencing, **36–37**

Roundworms (*Ascaris lumbricoides*), **81–82**

rRNA, *See* **Ribosomal RNA (rRNA)**

rRNA depletion method, **121–122**

RT, *See* **Retention time (RT)**

rUTIs, *See* **Recurrent urinary tract infections (rUTIs)**

S

16S rRNA gene, **16**, **80–81**, **153**, **165**

18S rRNA, **75–76**

16S sequencing, **153–154**

SAGs, *See* **Single-cell amplified genomes (SAGs)**

Salmonella

S-infected human cells, **133**

S. typhimurium, **118**

Sample acquisition and storage, **8–10**

Sample extraction, **231–232**

Sample handling considerations for metabolomics analyses, **227–229**

Sample processing, **56–57**

Sample-intrinsic microbial DNA found by tagging and sequencing (SIFT-seq), **56–57**

Sampling, **9–10**

and analysis pipelines, **39**

cost, **10**

Sanger sequencing, **32–33**, **34–36**

SBS, *See* **Sequencing by synthesis (SBS)**

Scaffolds, **14**

SCFAs, *See* **Short-chain fatty acids (SCFAs)**

SCGs, *See* **Single-copy genes (SCGs)**

Schistocerca gregaria, *See* **Desert Locust (*Schistocerca gregaria*)**

Scientific community, **197**

scRNA-seq, *See* **Single-cell RNA-seq (scRNA-seq)**

Secondary endosymbiosis, **73**

Sensitivity, **246–248**

Sequence data, **50**

Sequence-independent single-primer amplification, **157**

Sequencing

- coverage, **11–12**
- data, **166–167**
- devices, **146**
- platforms, **8**

Sequencing by synthesis (SBS), **34**

Serum, **228**

sgRNA, *See* **Single guide RNA (sgRNA)**

Shannon diversity

- index, **98–99**, **99*f***
- values, **99**

Shifting methodologies, **34–38**

Short-chain fatty acids (SCFAs), **243–244**

Short-read sequencing

platforms, **194–195**

technologies, **146**

Shotgun metagenomics, **168–171**

Shotgun sequencing, **37**

SIFT-seq, *See* **Sample-intrinsic microbial DNA found by tagging and sequencing (SIFT-seq)**

Similarity network fusion (SNF), **255–256**

Single guide RNA (sgRNA), **124–125**

Single-cell amplified genomes (SAGs), **196–197**

Single-cell RNA-seq (scRNA-seq), **134**

ribosomal RNA depletion in, **134–136**

Single-copy genes (SCGs), **175–176**

Single-molecule real-time (SMRT), **34**, **164**

Single-stranded DNA (ssDNA), **122**

Small molecules, **225–226**

smRandom-seq, **135**

SMRT, *See* **Single-molecule real-time (SMRT)**

SNF, *See* **Similarity network fusion (SNF)**

Soil microbiomes, **18**

Solid-phase extraction (SPE), **231**

Solvent-based extractions, **231**

Sparseness, **100**

SPE, *See* **Solid-phase extraction (SPE)**

Sphingolipids, **245–246**

ssDNA, *See* **Single-stranded DNA (ssDNA)**

Staphylococcus aureus, **60**, **121–122**

Statistical analysis

    of microbiome, **219**

    plan, **4–5**

Statistical hypothesis testing, **96–97**

    and design, **96–97**

Statistical model, **207–208**

Stool microbiome demonstrates, **220–221**

Storage, **9–10**

    conditions, **227–228**

Strain heterogeneity, **195**

Stremenopiles, **76–77**

Streptococcus, **100**, **103**

Structural variants (SVs), **148**

Subtractive hybridization, **115–121**, **117–118**, **124**

Sulfolobus solfataricus, **114–115**

Supervised methods, **255–256**

SVs, *See* **Structural variants (SVs)**

Symbiodinium, **73**

Symbioses, **67**

Page 289 choanoflagellates microbiome and origin of metazoans, **77–78**

endosymbionts with potential for denitrification and aerobic respiration, **75–76**

endosymbioses

and plastids evolution, **72–73**

in origin of eukaryotes and mitochondria, **71–72**

helminths, **81–85**

host genotype structures microbiome of diatom, **76–77**

microbiomes

and symbioses across invertebrates, **79–81**

within microbiomes, **81–85**

symbioses in protists, **73–75**

Symbiotic communities, **70**

T

t-tests, **100**

TAD, *See* **Typical American diet (TAD)**

Tamias amoenus cratericus, *See* **Chipmunk (*Tamias amoenus cratericus*)**

Targeted amplicon sequencing, **163–164**

Targeted analysis approach, **226–227**

Targeted metabolomics, **226–227**

Targeted metagenomics, **14**

Targeted molecular methods, **47**

Taxonomic precision, **165–166**

Taxonomic profiling, **210**

Teladorsagia circumcincta, **81–82**

Temporal limitations, **5**

Terminal restriction fragment length polymorphism (TRFLP), **32–33**

Terminator 5′-Phosphate-Dependent Exonuclease, **114–115**

Terrestrial soils, **244**

Tertiary endosymbiosis, **73**

Thalassiosira rotula, **77**

Third-generation sequencers, **34**, **59–60**

Threat-agnostic sequencing, **53**

Tiled amplicon sequencing, **55–56**

Tissue, **230**

    biopsies, **230**

TNFSF9, *See* **Tumor necrosis factor superfamily, member 9 (TNFSF9)**

Traditional microbiology, **29–30**

Traditional RNA-seq, **134**

Traditional technologies, **29–34**

Transcriptome, **2**, **109**, **129–130**

Transcriptomics, **2**, **14**, **109**

Transfer RNA (tRNA), **109**

TRFLP, *See* **Terminal restriction fragment length polymorphism (TRFLP)**

Trichoplax sp., **80**

Trichuris trichiura, *See* **Whipworms (*Trichuris trichiura*)**

5′-triphosphates, **114–115**

tRNA, *See* **Transfer RNA (tRNA)**

Tumor necrosis factor superfamily, member 9 (TNFSF9), **258–259**

Type 2 diabetes, **225**

Typical American diet (TAD), **98–99**

U

Untargeted metabolomics, **226–227**

Urine, **228–229**

metabolomes, **228–229**

US Department of Agriculture, **50–51**

US Department of Defense Global Emerging Infections Surveillance program, **51**

US Department of Health and Human Services, **50–51**

V

*VanR-O g*ene, **156–157**

Variance, **103–104**

Variation in biological systems, **39**

Versatility, **246–248**

Vertical integration, **252**

Vibrio, **49**

Virome, **157**

Viruses, **201**

W

Whipworms (*Trichuris trichiura*), **81–82**

Whole blood samples, **228**

Whole genome-shotgun sequencing, **1**

Wilcoxon test, **98–99**

Wolbachia, **83**

endosymbionts, **133**

World Health Organization (WHO), **6**

Wuchereria bancrofti, **83**

X

Xenobiotics, **16**

Y

Yersinia

Y. pestis, **53–55**

Y. pseudotuberculosis, **50–51**

Z

Zero-inflated beta regression model, **220**